Numerische Mathematik

Band 14 · (Schluß-) Heft 5 · 1970

Springer-Verlag · Berlin · Heidelberg · New York

Numer. Math. 10. IV. 1970

Die Zeitschrift „Numerische Mathematik" veröffentlicht auf breiter internationaler Grundlage Arbeiten, die sich mit allgemeinen Problemen des digitalen Rechnens, mit der Diskussion bestehender und der Entwicklung neuer numerischer Verfahren beschäftigen. Dabei werden die numerischen und programmierungstechnischen Gesichtspunkte des Einsatzes von Rechenautomaten im Vordergrund stehen.

The journal "Numerische Mathematik" provides for the international dissemination of contributions dealing with the general problems of digital computation. Such contributions may include discussions of existing numerical techniques as well as the development of new ones, but preferably with reference to the application of these techniques to programming for automatic computation.

Die Zeitschrift erscheint, um eine rasche Publikation zu ermöglichen, in einzelnen Heften, die zu Bänden vereinigt werden. Ein Band besteht im allgemeinen aus 5 Heften. Der Preis eines Bandes beträgt DM 96,—.

Grundsätzlich dürfen nur Arbeiten eingereicht werden, die vorher weder im Inland noch im Ausland veröffentlicht worden sind. Der Autor verpflichtet sich, sie auch nachträglich nicht an anderer Stelle zu publizieren. Mit der Annahme des Manuskriptes und seiner Veröffentlichung durch den Verlag geht auch das Recht der fotomechanischen Wiedergabe oder einer sonstigen Vervielfältigung an den Verlag über. Jedoch wird gewerblichen Unternehmen für den innerbetrieblichen Gebrauch nach Maßgabe des zwischen dem Börsenverein des Deutschen Buchhandels e. V. und dem Bundesverband der Deutschen Industrie abgeschlossenen Rahmenabkommens die Anfertigung einer fotomechanischen Vervielfältigung gestattet. Wenn für diese Zeitschrift kein Pauschalabkommen mit dem Verlag vereinbart ist, ist eine Wertmarke im Betrage von DM 0,30 pro Seite zu verwenden. *Der Verlag läßt diese Beträge den Autorenverbänden zufließen.*

Die Mitarbeiter erhalten von ihren Arbeiten zusammen 75 Sonderdrucke unentgeltlich.

Manuskriptsendungen und Anfragen können an jeden der Herausgeber gerichtet werden (Anschriften s. 3. Umschlagseite).

Bei der Anlage der Manuskripte sind die „Hinweise für die Autoren" am Schluß des Heftes zu beachten.

In the interest of speedy publication this journal is issued at frequent intervals according to the material received. As a rule 5 numbers constitute one volume. The price is DM 96.— per volume.

No paper will be acceptable that has had previous publication either here or abroad, and, moreover, it is understood that if accepted, the author will not have it published elsewhere subsequently. The Publisher acquires the rights of photographic and other forms of reproduction, upon acceptance and publication of a manuscript.

The author or the authors of a given paper will receive a total of 75 reprints at no charge.

Manuscripts and enquiries may be directed to any member of the editorial board ("Herausgeber") listed on the third cover page.

When preparing the manuscripts please consider the "Directions for Authors" at the end of the issue.

Springer-Verlag

69 Heidelberg 1
Postfach 1780
Fernsprecher (06221) 49101
Fernschreibnummer 04-61723

1 Berlin 33
Heidelberger Platz 3
Fernsprecher (0311) 822001
Fernschreibnummer 01-83319

Springer-Verlag
New York Inc.
175 Fifth Avenue
New York, N. Y. 10010

Numer. Math. 14, 403–420 (1970)

Handbook Series Linear Algebra

Singular Value Decomposition and Least Squares Solutions*

Contributed by

G. H. GOLUB** and C. REINSCH

1. Theoretical Background

1.1. Introduction

Let A be a real $m \times n$ matrix with $m \geqq n$. It is well known (cf. [4]) that

$$A = U \Sigma V^T \tag{1}$$

where

$$U^T U = V^T V = V V^T = I_n \quad \text{and} \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n).$$

The matrix U consists of n orthonormalized eigenvectors associated with the n largest eigenvalues of $A A^T$, and the matrix V consists of the orthonormalized eigenvectors of $A^T A$. The diagonal elements of Σ are the non-negative square roots of the eigenvalues of $A^T A$; they are called *singular values*. We shall assume that

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n \geqq 0.$$

Thus if $rank(A) = r$, $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$. The decomposition (1) is called the *singular value decomposition* (SVD).

There are alternative representations to that given by (1). We may write

$$A = U_c \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T \quad \text{with} \quad U_c^T U_c = I_m$$

or

$$A = U_r \Sigma_r V_r^T \quad \text{with} \quad U_r^T U_r = V_r^T V_r = I_r \quad \text{and} \quad \Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r).$$

We use the form (1), however, since it is most useful for computational purposes.

If the matrix U is not needed, it would appear that one could apply the usual diagonalization algorithms to the symmetric matrix $A^T A$ which has to be formed explicitly. However, as in the case of linear least squares problems, the com-

* *Editor's note.* In this fascicle, prepublication of algorithms from the Linear Algebra series of the Handbook for Automatic Computation is continued. Algorithms are published in ALGOL 60 reference language as approved by the IFIP. Contributions in this series should be styled after the most recently published ones.

** The work of this author was in part supported by the National Science Foundation and Office of Naval Research.

putation of $A^T A$ involves unnecessary numerical inaccuracy. For example, let

$$A = \begin{bmatrix} 1 & 1 \\ \beta & 0 \\ 0 & \beta \end{bmatrix},$$

then

$$A^T A = \begin{bmatrix} 1+\beta^2 & 1 \\ 1 & 1+\beta^2 \end{bmatrix}$$

so that

$$\sigma_1(A) = (2+\beta^2)^{\frac{1}{2}}, \quad \sigma_2(A) = |\beta|.$$

If $\beta^2 < \varepsilon_0$, the machine precision, the computed $A^T A$ has the form $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, and the best one may obtain from diagonalization is $\tilde{\sigma}_1 = \sqrt{2}$, $\tilde{\sigma}_2 = 0$.

To compute the singular value decomposition of a given matrix A, Forsythe and Henrici [2], Hestenes [8], and Kogbetliantz [9] proposed methods based on plane rotations. Kublanovskaya [10] suggested a QR-type method. The program described below first uses Householder transformations to reduce A to bidiagonal form, and then the QR algorithm to find the singular values of the bidiagonal matrix. The two phases properly combined produce the singular value decomposition of A.

1.2. Reduction to Bidiagonal Form

It was shown in [6] how to construct two finite sequences of Householder transformations

$$P^{(k)} = I - 2x^{(k)} x^{(k)T} \quad (k = 1, 2, \ldots, n)$$

and

$$Q^{(k)} = I - 2y^{(k)} y^{(k)T} \quad (k = 1, 2, \ldots, n-2)$$

(where $x^{(k)T} x^{(k)} = y^{(k)T} y^{(k)} = 1$) such that

$$P^{(n)} \ldots P^{(1)} A Q^{(1)} \ldots Q^{(n-2)} = \begin{bmatrix} q_1 & e_2 & 0 & \cdots & 0 \\ & q_2 & e_3 & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ & & & \ddots & e_n \\ & 0 & & & q_n \\ \hline & & 0 & & \end{bmatrix} \equiv J^{(0)},$$

$\left.\begin{bmatrix} \\ \\ \\ \\ \\ \hline \\ \end{bmatrix}\right\} (m-n) \times n$

an upper bidiagonal matrix. If we let $A^{(1)} = A$ and define

$$A^{(k+\frac{1}{2})} = P^{(k)} A^{(k)} \quad (k = 1, 2, \ldots, n)$$

$$A^{(k+1)} = A^{(k+\frac{1}{2})} Q^{(k)} \quad (k = 1, 2, \ldots, n-2)$$

then $P^{(k)}$ is determined such that

$$a_{ik}^{(k+\frac{1}{2})} = 0 \quad (i = k+1, \ldots, m)$$

and $Q^{(k)}$ such that

$$a_{kj}^{(k+1)} = 0 \quad (j = k+2, \ldots, n).$$

The singular values of $J^{(0)}$ are the same as those of A. Thus, if the singular value decomposition of

$$J^{(0)} = G \Sigma H^T$$

then

$$A = PG \Sigma H^T Q^T$$

so that $U = PG$, $V = QH$ with $P \equiv P^{(1)} \dots P^{(n)}$, $Q \equiv Q^{(1)} \dots Q^{(n-2)}$.

1.3. Singular Value Decomposition of the Bidiagonal Matrix

By a variant of the QR algorithm, the matrix $J^{(0)}$ is iteratively diagonalized so that

$$J^{(0)} \to J^{(1)} \to \cdots \to \Sigma$$

where

$$J^{(i+1)} = S^{(i)T} J^{(i)} T^{(i)},$$

and $S^{(i)}$, $T^{(i)}$ are orthogonal. The matrices $T^{(i)}$ are chosen so that the sequence $M^{(i)} = J^{(i)T} J^{(i)}$ converges to a diagonal matrix while the matrices $S^{(i)}$ are chosen so that all $J^{(i)}$ are of the bidiagonal form. In [7], another technique for deriving $\{S^{(i)}\}$ and $\{T^{(i)}\}$ is given but this is equivalent to the method described below.

For notational convenience, we drop the suffix and use the notation

$$J \equiv J^{(i)}, \quad \bar{J} \equiv J^{(i+1)}, \quad S \equiv S^{(i)}, \quad T \equiv T^{(i)}, \quad M \equiv J^T J, \quad \bar{M} \equiv \bar{J}^T \bar{J}.$$

The transition $J \to \bar{J}$ is achieved by application of Givens rotations to J alternately from the right and the left. Thus

$$\bar{J} = \underbrace{S_n^T S_{(n-1)}^T \dots S_2^T}_{S^T} J \underbrace{T_2 T_3 \dots T_n}_{T} \tag{2}$$

where

$$S_k = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & 0 & \\ & & 1 & & & & & & \\ & & & \cos\theta_k & -\sin\theta_k & & & & \\ & & & \sin\theta_k & \cos\theta_k & & & & \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & 0 & & & & & 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ (k-1) \\ (k) \\ \\ \\ \\ \end{matrix}$$

and T_k is defined analogously to S_k with φ_k instead of θ_k.

Let the first angle, φ_2, be arbitrary while all the other angles are chosen so that $\bar{J}$ has the same form as J. Thus,

$$\begin{aligned} &T_2 \text{ annihilates nothing,} && \text{generates an entry } \{J\}_{21}, \\ &S_2^T \text{ annihilates } \{J\}_{21}, && \text{generates an entry } \{J\}_{13}, \\ &T_3 \text{ annihilates } \{J\}_{13}, && \text{generates an entry } \{J\}_{32}, \\ &\quad\vdots \end{aligned} \tag{3}$$

and finally

S_n^T annihilates $\{J\}_{n,n-1}$, and generates nothing.

(See Fig. 1.)

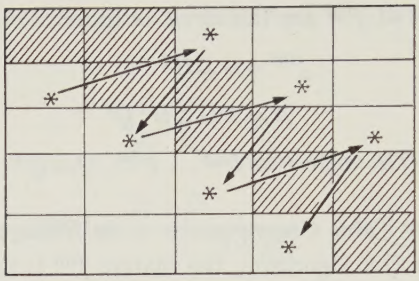

Fig. 1

This process is frequently described as "chasing". Since $\bar{J} = S^T J T$,

$$\overline{M} = \bar{J}^T \bar{J} = T^T M T$$

and $\overline{M}$ is a tri-diagonal matrix just as M is. We show that the first angle, φ_2, which is still undetermined, can be chosen so that the transition $M \to \overline{M}$ is a QR transformation with a given shift s.

The usual QR algorithm with shifts is described as follows:

$$\begin{aligned} M - sI &= T_s R_s \\ R_s T_s + sI &= \overline{M}_s \end{aligned} \tag{4}$$

where $T_s^T T_s = I$ and R_s is an upper triangular matrix. Thus $\overline{M}_s = T_s^T M T_s$. It has been shown by Francis [5] that it is not necessary to compute (4) explicitly but it is possible to perform the shift implicitly. Let T be for the moment an arbitrary matrix such that

$$\{T_s\}_{k,1} = \{T\}_{k,1} \quad (k = 1, 2, \ldots, n),$$

(i.e., the elements of the first column of T_s are equal to the first column of T) and

$$T^T T = I.$$

Then we have the following theorem (Francis): If

i) $\overline{M} = T^T M T$,

ii) $\overline{M}$ is a tri-diagonal matrix,

iii) the sub-diagonal elements of M are non-zero,

it follows that $\overline{M} = D \overline{M}_s D$ where D is a diagonal matrix whose diagonal elements are ± 1.

Thus choosing T_2 in (3) such that its first column is proportional to that of $M - sI$, the same is true for the first column of the product $\dot{T} = T_2 T_3 \ldots T_n$ which therefore is identical to that of T_s. Hence, if the sub-diagonal of M does not contain any non-zero entry the conditions of the Francis theorem are fulfilled and T is therefore identical to T_s (up to a scaling of column ± 1). Thus the transition (2) is equivalent to the QR transformation of $J^T J$ with a given shift s.

The shift parameter s is determined by an eigenvalue of the lower 2×2 minor of M. Wilkinson [13] has shown that for this choice of s, the method converges globally and almost always cubically.

1.4. Test for Convergence

If $|e_n| \leqq \delta$, a prescribed tolerance, then $|q_n|$ is accepted as a singular value, and the order of the matrix is dropped by one. If, however, $|e_k| \leqq \delta$ for $k \neq n$, the matrix breaks into two, and the singular values of each block may be computed independently.

If $q_k = 0$, then at least one singular value must be equal to zero. In the absence of roundoff error, the matrix will break if a shift of zero is performed. Now, suppose at some stage

$$|q_k| \leqq \delta.$$

At this stage an extra sequence of Givens rotations is applied from the left to J involving rows $(k, k+1)$, $(k, k+2)$, ..., (k, n) so that

$e_{k+1} \equiv \{J\}_{k,k+1}$ is annihilated, but $\{J\}_{k,k+2}$, $\{J\}_{k+1,k}$ are generated,
$\{J\}_{k,k+2}$ is annihilated, but $\{J\}_{k,k+3}$, $\{J\}_{k+2,k}$ are generated,
$\vdots$
and finally $\{J\}_{k,n}$ is annihilated, and $\{J\}_{n,k}$ is generated.

The matrix obtained thusly has the form

$$\bar{J} = \begin{bmatrix} & & & & & & \\ & & & & & 0 & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & 0 & & & & & \\ & & & & & & \end{bmatrix}^{(k)} \begin{matrix} q_1 & e_2 & & & & & \\ & \ddots & \ddots & & & & \\ & & \ddots & e_k & & & \\ & & & \bar{q}_k & 0 & & \\ & & & \delta_{k+1} & \bar{q}_{k+1} & \bar{e}_{k+2} & \\ & & & \vdots & & \ddots & \ddots \\ & & & \delta_n & & & \bar{e}_n \\ & & & & & & \bar{q}_n \end{matrix}^{(k)}.$$

Note that by orthogonality

$$\bar{q}_k^2 + \delta_{k+1}^2 + \cdots + \delta_n^2 = q_k^2 \leqq \delta^2.$$

Thus choosing $\delta = \|J^{(0)}\|_\infty \varepsilon_0$ (ε_0, the machine precision) ensures that all δ_k are less in magnitude than $\varepsilon_0 \|J^{(0)}\|_\infty$. Elements of $\bar{J}$ not greater than this are neglected. Hence $\bar{J}$ breaks up into two parts which may be treated independently.

2. Applicability

There are a large number of applications of the singular value decomposition; an extensive list is given in [7]. Some of these are as follows:

2.1. Pseudoinverse (Procedure SVD)

Let A be a real $m \times n$ matrix. An $n \times m$ matrix X is said to be the pseudoinverse of A if X satisfies the following four properties:

i) $AXA = A$,
ii) $XAX = X$,
iii) $(AX)^T = AX$,
iv) $(XA)^T = XA$.

The unique solution is denoted by A^+. It is easy to verify that if $A = U\Sigma V^T$, then $A^+ = V\Sigma^+ U^T$ where $\Sigma^+ = \text{diag}(\sigma_i^+)$ and

$$\sigma_i^+ = \begin{cases} 1/\sigma_i & \text{for} \quad \sigma_i > 0 \\ 0 & \text{for} \quad \sigma_i = 0. \end{cases}$$

Thus the pseudoinverse may easily be computed from the output provided by the procedure *SVD*.

2.2. Solution of Homogeneous Equations (Procedure SVD or Procedure Minfit)

Let A be a matrix of rank r, and suppose we wish to solve

$$A x_i = \theta \quad \text{for} \quad i = r+1, \ldots, n$$

where θ denotes the null vector.

Let

$$U = [u_1, u_2, \ldots, u_n] \quad \text{and} \quad V = [v_1, v_2, \ldots, v_n].$$

Then since $A v_i = \sigma_i u_i$ $(i = 1, 2, \ldots n)$,

$$A v_i = \theta \quad \text{for} \quad i = r+1, \ldots, n$$

and $x_i = v_i$.

Here the procedure *SVD* or the procedure *Minfit* with $p = 0$ may be used for determining the solution. If the rank of A is known, then a modification of the algorithm of Businger and Golub [1] may be used.

2.3. Solutions of Minimal Length (Procedure Minfit)

Let b_1 be a given vector. Suppose we wish to determine a vector x so that

$$\|b_1 - A x\|_2 = \min \tag{5}$$

If the rank of A is less than n then there is no unique solution. Thus we require amongst all x which satisfy (5) that

$$\|\hat{x}\|_2 = \min$$

and this solution is unique. It is easy to verify that

$$\hat{x} = A^+ b_1 = V\Sigma^+ U^T b_1 \equiv V\Sigma^+ c_1.$$

The procedure *Minfit* with $p > 0$ will yield V, Σ, $c_1, \ldots, c_p$. Thus the user is able to determine which singular values are to be declared as zero.

2.4. A Generalization of the Least Squares Problem (Procedure SVD)

Let A be a real $m \times n$ matrix of rank n and let b be a given vector. We wish to construct a vector x such that

$$(A + \Delta A)\, x = b + \Delta b$$

and

$$\text{trace}(\Delta A^T \Delta A) + K^2 \Delta b^T \Delta b = \min.$$

Here $K>0$ is a given weight and the standard problem is obtained for $K\to 0$. Introducing the augmented matrices $\bar{A}=(A, Kb)$ and $\Delta\bar{A}=(\Delta A, K\Delta b)$ and the vector

$$\bar{x}=\begin{pmatrix} x \\ -1/K \end{pmatrix},$$

we have to minimize $\operatorname{trace}(\Delta\bar{A}^T\Delta\bar{A})$ under the constraint $(\bar{A}+\Delta\bar{A})\bar{x}=0$. For fixed $\bar{x}$ the minimum is attained for $\Delta\bar{A}=-\bar{A}\bar{x}\bar{x}^T/\bar{x}^T\bar{x}$ and it has the value $\bar{x}^T\bar{A}^T\bar{A}\bar{x}/\bar{x}^T\bar{x}$. Minimizing with respect to $\bar{x}$ amounts to the computation of the smallest singular value of the matrix $\bar{A}$ and $\bar{x}$ is the corresponding column of the matrix $\bar{V}$ in the decomposition (1) with proper normalization [3].

3. Formal Parameter List

3.1. Input to Procedure SVD

m — number of rows of A, $m\geqq n$.

n — number of columns of A.

withu — **true** if U is desired, **false** otherwise.

withv — **true** if V is desired, **false** otherwise.

eps — a constant used in the test for convergence (see Section 5, (iii)); should not be smaller than the machine precision ε_0, i.e., the smallest number for which $1+\varepsilon_0>1$ in computer arithmetic.

tol — a machine dependent constant which should be set equal to β/ε_0 where β is the smallest positive number representable in the computer, see [11].

a[1:*m*, 1:*n*] — represents the matrix A to be decomposed.

Output of procedure *SVD*.

q[1:*n*] — a vector holding the singular values of A; they are non-negative but not necessarily ordered in decreasing sequence.

u[1:*m*, 1:*n*] — represents the matrix U with orthonormalized columns (if *withu* is **true**, otherwise *u* is used as a working storage).

v[1:*n*, 1:*n*] — represents the orthogonal matrix V (if *withv* is **true**, otherwise *v* is not used).

3.2. Input to Procedure Minfit

m — number of rows of A.

n — number of columns of A.

p — number of columns of B, $p\geqq 0$.

eps — same as for procedure *SVD*.

tol — same as for procedure *SVD*.

ab[1:*max*(*m*, *n*), 1:*n*+*p*] — *ab*[*i*, *j*] represents $a_{i,j}$, $1\leqq i\leqq m$, $1\leqq j\leqq n$,
ab[*i*, *n*+*j*] represents $b_{i,j}$, $1\leqq i\leqq m$, $1\leqq j\leqq p$.

Output of procedure *Minfit*.

$ab[1:max(m,n), 1:n+p]$	$ab[i, j]$ represents $v_{i,j}$, $1 \leqq i \leqq n$, $1 \leqq j \leqq n$, $ab[i, n+j]$ represents $c_{i,j}$, $1 \leqq i \leqq max(m,n)$, $1 \leqq j \leqq p$ viz. $C = U_c^T B$.
$q[1:n]$	same as for procedure *SVD*.

4. ALGOL Programs

```
procedure SVD (m, n, withu, withv, eps, tol) data: (a) result: (q, u, v);
    value m, n, withu, withv, eps, tol;
    integer m, n;
    Boolean withu, withv;
    real eps, tol;
    array a, q, u, v;
```

comment Computation of the singular values and complete orthogonal decomposition of a real rectangular matrix A,

$$A = U \operatorname{diag}(q)\, V^T, \qquad U^T U = V^T V = I,$$

where the arrays $a[1:m, 1:n]$, $u[1:m, 1:n]$, $v[1:n, 1:n]$, $q[1:n]$ represent A, U, V, q respectively. The actual parameters corresponding to a, u, v may all be identical unless *withu* = *withv* = **true**. In this case, the actual parameters corresponding to u and v must differ. $m \geqq n$ is assumed;

```
begin
    integer i, j, k, l, l1;
    real c, f, g, h, s, x, y, z;
    array e[1:n];
    for i := 1 step 1 until m do
    for j := 1 step 1 until n do u[i, j] := a[i, j];
```

comment Householder's reduction to bidiagonal form;

```
    g := x := 0;
    for i := 1 step 1 until n do
    begin
        e[i] := g; s := 0; l := i+1;
        for j := i step 1 until m do s := s+u[j, i]↑2;
        if s<tol then g := 0 else
        begin
            f := u[i, i]; g := if f<0 then sqrt(s) else -sqrt(s);
            h := f×g-s; u[i, i] := f-g;
            for j := l step 1 until n do
            begin
                s := 0;
                for k := i step 1 until m do s := s+u[k, i]×u[k, j];
                f := s/h;
                for k := i step 1 until m do u[k, j] := u[k, j]+f×u[k, i]
            end j
        end s;
```

```
        q[i] := g; s := 0;
        for j := l step 1 until n do s := s + u[i, j]↑2;
        if s < tol then g := 0 else
        begin
            f := u[i, i+1]; g := if f < 0 then sqrt(s) else -sqrt(s);
            h := f × g - s; u[i, i+1] := f - g;
            for j := l step 1 until n do e[j] := u[i, j]/h;
            for j := l step 1 until m do
            begin
                s := 0;
                for k := l step 1 until n do s := s + u[j, k] × u[i, k];
                for k := l step 1 until n do u[j, k] := u[j, k] + s × e[k]
            end j
        end s;
        y := abs(q[i]) + abs(e[i]); if y > x then x := y
    end i;

comment accumulation of right-hand transformations;
    if withv then for i := n step -1 until 1 do
    begin
        if g ≠ 0 then
        begin
            h := u[i, i+1] × g;
            for j := l step 1 until n do v[j, i] := u[i, j]/h;
            for j := l step 1 until n do
            begin
                s := 0;
                for k := l step 1 until n do s := s + u[i, k] × v[k, j];
                for k := l step 1 until n do v[k, j] := v[k, j] + s × v[k, i]
            end j
        end g;
        for j := l step 1 until n do v[i, j] := v[j, i] := 0;
        v[i, i] := 1; g := e[i]; l := i
    end i;

comment accumulation of left-hand transformations;
    if withu then for i := n step -1 until 1 do
    begin
        l := i + 1; g := q[i];
        for j := l step 1 until n do u[i, j] := 0;
        if g ≠ 0 then
        begin
            h := u[i, i] × g;
            for j := l step 1 until n do
```

```
            begin
                s := 0;
                for k := l step 1 until m do s := s + u[k, i] × u[k, j];
                f := s/h;
                for k := i step 1 until m do u[k, j] := u[k, j] + f × u[k, i]
            end j;
            for j := i step 1 until m do u[j, i] := u[j, i]/g
        end g
        else for j := i step 1 until m do u[j, i] := 0;
        u[i, i] := u[i, i] + 1
    end i;

comment diagonalization of the bidiagonal form;
    eps := eps × x;
    for k := n step −1 until 1 do
    begin
        test f splitting:
            for l := k step −1 until 1 do
            begin
                if abs(e[l]) ≦ eps then goto test f convergence;
                if abs(q[l − 1]) ≦ eps then goto cancellation
            end l;

comment cancellation of e[l] if l > 1;
        cancellation:
            c := 0; s := 1; l1 := l − 1;
            for i := l step 1 until k do
            begin
                f := s × e[i]; e[i] := c × e[i];
                if abs(f) ≦ eps then goto test f convergence;
                g := q[i]; h := q[i] := sqrt(f × f + g × g); c := g/h; s := −f/h;
                if withu then for j := 1 step 1 until m do
                begin
                    y := u[j, l1]; z := u[j, i];
                    u[j, l1] := y × c + z × s; u[j, i] := −y × s + z × c
                end j
            end i;
        test f convergence:
            z := q[k]; if l = k then goto convergence;

comment shift from bottom 2 × 2 minor;
            x := q[l]; y := q[k − 1]; g := e[k − 1]; h := e[k];
            f := ((y − z) × (y + z) + (g − h) × (g + h))/(2 × h × y); g := sqrt(f × f + 1);
            f := ((x − z) × (x + z) + h × (y/(if f < 0 then f − g else f + g) − h))/x;

comment next Q R transformation;
            c := s := 1;
            for i := l + 1 step 1 until k do
```

```
        begin
            g := e[i]; y := q[i]; h := s×g; g := c×g;
            e[i−1] := z := sqrt(f×f+h×h); c := f/z; s := h/z;
            f := x×c+g×s; g := −x×s+g×c; h := y×s; y := y×c;
            if withv then for j := 1 step 1 until n do
            begin
                x := v[j, i−1]; z := v[j, i];
                v[j, i−1] := x×c+z×s; v[j, i] := −x×s+z×c
            end j;
            q[i−1] := z := sqrt(f×f+h×h); c := f/z; s := h/z;
            f := c×g+s×y; x := −s×g+c×y;
            if withu then for j := 1 step 1 until m do
            begin
                y := u[j, i−1]; z := u[j, i];
                u[j, i−1] := y×c+z×s; u[j, i] := −y×s+z×c
            end j
        end i;
        e[l] := 0; e[k] := f; q[k] := x; goto test f splitting;
    convergence:
        if z<0 then
        begin comment q[k] is made non-negative;
            q[k] := −z;
            if withv then for j := 1 step 1 until n do v[j, k] := −v[j, k]
        end z
    end k
end SVD;


procedure Minfit (m, n, p, eps, tol) trans: (ab) result: (q);
    value m, n, p, eps, tol;
    integer m, n, p;
    real eps, tol;
    array ab, q;
```

comment Computation of the matrices diag(q), V, and C such that for given real $m\times n$ matrix A and $m\times p$ matrix B

$$U_c^T A V = \text{diag}(q) \text{ and } U_c^T B = C \text{ with orthogonal matrices } U_c \text{ and } V.$$

The singular values and the matrices V and C may be used to determine $\bar{X}$ minimizing (1) $\|AX-B\|_F$ and (2) $\|X\|_F$ with the solution

$$\bar{X} = V \times \text{Pseudo-inverse of diag}(q) \times C.$$

The procedure can also be used to determine the complete solution of an underdetermined linear system, i.e., rank$(A)=m<n$.

The array $q[1:n]$ represents the matrix diag(q). A and B together are to be given as the first m rows of the array $ab[1:max(m,n), 1:n+p]$. V is returned in the first n rows and columns of ab while C is returned in the last p columns of ab (if $p>0$);

```
begin
    integer i, j, k, l, l1, n1, np;
    real c, f, g, h, s, x, y, z;
    array e[1:n];

comment Householder's reduction to bidiagonal form;
    g := x := 0; np := n + p;
    for i := 1 step 1 until n do
    begin
        e[i] := g; s := 0; l := i + 1;
        for j := i step 1 until m do s := s + ab[j, i]↑2;
        if s < tol then g := 0 else
        begin
            f := ab[i, i]; g := if f < 0 then sqrt(s) else -sqrt(s);
            h := f × g - s; ab[i, i] := f - g;
            for j := l step 1 until np do
            begin
                s := 0;
                for k := i step 1 until m do s := s + ab[k, i] × ab[k, j];
                f := s/h;
                for k := i step 1 until m do ab[k, j] := ab[k, j] + f × ab[k, i]
            end j
        end s;
        q[i] := g; s := 0;
        if i ≦ m then for j := l step 1 until n do s := s + ab[i, j]↑2;
        if s < tol then g := 0 else
        begin
            f := ab[i, i + 1]; g := if f < 0 then sqrt(s) else -sqrt(s);
            h := f × g - s; ab[i, i + 1] := f - g;
            for j := l step 1 until n do e[j] := ab[i, j]/h;
            for j := l step 1 until m do
            begin
                s := 0;
                for k := l step 1 until n do s := s + ab[j, k] × ab[i, k];
                for k := l step 1 until n do ab[j, k] := ab[j, k] + s × e[k]
            end j
        end s;
        y := abs(q[i]) + abs(e[i]); if y > x then x := y
    end i;

comment accumulation of right-hand transformations;
    for i := n step -1 until 1 do
    begin
        if g ≠ 0 then
        begin
            h := ab[i, i + 1] × g;
            for j := l step 1 until n do ab[j, i] := ab[i, j]/h;
            for j := l step 1 until n do
```

```
            begin
                s := 0;
                for k := l step 1 until n do s := s + ab[i, k] × ab[k, j];
                for k := l step 1 until n do ab[k, j] := ab[k, j] + s × ab[k, i]
            end j
        end g;
        for j := l step 1 until n do ab[i, j] := ab[j, i] := 0;
        ab[i, i] := 1; g := e[i]; l := i
    end i;
    eps := eps × x; n1 := n + 1;
    for i := m + 1 step 1 until n do
    for j := n1 step 1 until np do ab[i, j] := 0;

comment diagonalization of the bidiagonal form;
    for k := n step −1 until 1 do
    begin
        test f splitting:
            for l := k step −1 until 1 do
            begin
                if abs(e[l]) ≦ eps then goto test f convergence;
                if abs(q[l − 1]) ≦ eps then goto cancellation
            end l;

comment cancellation of e[l] if l > 1;
        cancellation:
            c := 0; s := 1; l1 := l − 1;
            for i := l step 1 until k do
            begin
                f := s × e[i]; e[i] := c × e[i];
                if abs(f) ≦ eps then goto test f convergence;
                g := q[i]; q[i] := h := sqrt(f × f + g × g); c := g/h; s := −f/h;
                for j := n1 step 1 until np do
                begin
                    y := ab[l1, j]; z := ab[i, j];
                    ab[l1, j] := c × y + s × z; ab[i, j] := −s × y + c × z
                end j
            end i;
        test f convergence:
            z := q[k]; if l = k then goto convergence;

comment shift from bottom 2 × 2 minor;
            x := q[l]; y := q[k − 1]; g := e[k − 1]; h := e[k];
            f := ((y − z) × (y + z) + (g − h) × (g + h))/(2 × h × y); g := sqrt(f × f + 1);
            f := ((x − z) × (x + z) + h × (y/(if f < 0 then f − g else f + g) − h))/x;

comment next Q R transformation;
            c := s := 1;
            for i := l + 1 step 1 until k do
```

```
        begin
            g := e[i]; y := q[i]; h := s×g; g := c×g;
            e[i−1] := z := sqrt(f×f+h×h); c := f/z; s := h/z;
            f := x×c+g×s; g := −x×s+g×c; h := y×s; y := y×c;
            for j := 1 step 1 until n do
            begin
                x := ab[j, i−1]; z := ab[j, i];
                ab[j, i−1] := x×c+z×s; ab[j, i] := −x×s+z×c
            end j;
            q[i−1] := z := sqrt(f×f+h×h); c := f/z; s := h/z;
            f := c×g+s×y; x := −s×g+c×y;
            for j := n1 step 1 until np do
            begin
                y := ab[i−1, j]; z := ab[i, j];
                ab[i−1, j] := c×y+s×z; ab[i, j] := −s×y+c×z
            end j
        end i;
        e[l] := 0; e[k] := f; q[k] := x; goto test f splitting;
    convergence:
        if z<0 then
        begin comment q[k] is made non-negative;
            q[k] := −z;
            for j := 1 step 1 until n do ab[j, k] := −ab[j, k]
        end z
    end k
end Minfit;
```

5. Organizational and Notational Details

(i) The matrix U consists of the first n columns of an orthogonal matrix U_c. The following modification of the procedure *SVD* would produce U_c instead of U: After

```
comment accumulation of left-hand transformations;
```

insert a statement

```
if withu then for i := n+1 step 1 until m do
begin
    for j := n+1 step 1 until m do u[i, j] := 0;
    u[i, i] := 1
end i;
```

Moreover, replace n by m in the fourth and eighth line after that, i.e., write twice **for** $j := l$ **step** 1 **until** m **do**.

(ii) $m \geqq n$ is assumed for procedure *SVD*. This is no restriction; if $m < n$, store A^T, i.e., use an array $at[1:n, 1:m]$ where $at[i, j]$ represents $a_{j,i}$ and call *SVD*(*n, m, withv, withu, eps, tol, at, q, v, u*) producing the $m \times m$ matrix U and the $n \times m$ matrix V. There is no restriction on the values of m and n for the procedure *Minfit*.

(iii) In the iterative part of the procedures an element of $J^{(i)}$ is considered to be negligible and is consequently replaced by zero if it is not larger in magnitude than εx where ε is the given tolerance and

$$x = \max_{1 \leq i \leq n} (|q_i| + |e_i|).$$

The largest singular value σ_1 is bounded by $x/\sqrt{2} \leq \sigma_1 \leq x\sqrt{2}$.

(iv) A program organization was chosen which allows us to save storage locations. To this end the actual parameters corresponding to a and u may be identical. In this event the original information stored in a is overwritten by information on the reduction. This, in turn, is overwritten by u if the latter is desired. Likewise, the actual parameters corresponding to a and v may agree. Then v is stored in the upper part of a if it is desired, otherwise a is not changed. Finally, all three parameters a, u, and v may be identical unless *withu* = *withv* = **true**.

This special feature, however, increases the number of multiplications needed to form U roughly by a factor m/n.

(v) Shifts are evaluated in a way as to reduce the danger of overflow or underflow of exponents.

(vi) The singular values as delivered in the array q are not necessarily ordered. Any sorting of them should be accompanied by the corresponding sorting of the columns of U and V, and of the rows of C.

(vii) The formal parameter list may be completed by the addition of a limit for the number of iterations to be performed, and by the addition of a failure exit to be taken if no convergence is reached after the specified number of iterations (e.g., 30 per singular value).

6. Numerical Properties

The stability of the Householder transformations has been demonstrated by Wilkinson [12]. In addition, he has shown that in the absence of roundoff the QR algorithm has global convergence and asymptotically is almost always cubically convergent.

The numerical experiments indicate that the average number of complete QR iterations on the bidiagonal matrix is usually less than two per singular value. Extra consideration must be given to the implicit shift technique which fails for a split matrix. The difficulties arise when there are small q_k's or e_k's. Using the techniques of Section 1.4, there cannot be numerical instability since stable orthogonal transformations are used but under special circumstances there may be a slowdown in the rate of convergence.

7. Test Results

Tests were carried out on the UNIVAC 1108 Computer of the Andrew R. Jennings Computing Center of Case Western Reserve University. Floating point numbers are represented by a normalized 27 bit mantissa and a 7 bit exponent to the radix 2, whence $eps = 1.5_{10} - 8$, $tol = {}_{10} - 31$. In the following, computed values are marked by a tilde and $m(A)$ denotes $\max |a_{i,j}|$.

First example:

$$A = \begin{bmatrix} 22 & 10 & 2 & 3 & 7 \\ 14 & 7 & 10 & 0 & 8 \\ -1 & 13 & -1 & -11 & 3 \\ -3 & -2 & 13 & -2 & 4 \\ 9 & 8 & 1 & -2 & 4 \\ 9 & 1 & -7 & 5 & -1 \\ 2 & -6 & 6 & 5 & 1 \\ 4 & 5 & 0 & -2 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} -1 & 1 & 0 \\ 2 & -1 & 1 \\ 1 & 10 & 11 \\ 4 & 0 & 4 \\ 0 & -6 & -6 \\ -3 & 6 & 3 \\ 1 & 11 & 12 \\ 0 & -5 & -5 \end{bmatrix},$$

$$\sigma_1 = \sqrt{1248}, \quad \sigma_2 = 20, \quad \sigma_3 = \sqrt{384}, \quad \sigma_4 = \sigma_5 = 0.$$

The homogeneous system $A x = \theta$ has two linearly independent solutions. Six QR transformations were necessary to drop all off-diagonal elements below the internal tolerance $46.4_{10} - 8$. Table 1 gives the singular values in the sequence as computed by procedures *SVD* and *Minfit*. The accuracy of the achieved decomposition is characterized by

$$m(A - \tilde{U}\tilde{\Sigma}\tilde{V}^T) = 238_{10} - 8, \quad m(\tilde{U}^T\tilde{U} - I) = 8.1_{10} - 8, \quad m(\tilde{V}^T\tilde{V} - I) = 3.3_{10} - 8.$$

Because two singular values are equal to zero, the procedures *SVD* and *Minfit* may lead to other orderings of the singular values for this matrix when other tolerances are used.

Table 1

$\tilde{\sigma}_k$	$\sigma_k - \tilde{\sigma}_k$
$0.96_{10} - 7$	-9.6
19.595 916	191
19.999 999	143
$1.97_{10} - 7$	-19.7
35.327 038	518

$\times 10^{-8}$

The computed solutions of the homogeneous system are given by the first and fourth columns of the matrix $\tilde{V}$ (Table 2).

Table 2

$\tilde{v}_1$	$\tilde{v}_4$	$v_1 - \tilde{v}_1$	$v_4 - \tilde{v}_4$
$-0.4190\,9545$	0	-1.5	0 (Def.)
0.4405 0912	0.4185 4806	1.7	0.6
$-0.0520\,0457$	0.3487 9006	1.2	-1.3
0.6760 5915	0.2441 5305	1.0	0.3
0.4129 7730	$-0.8022\,1713$	1.3	-0.8

$\times 10^{-8}$

Procedure *Minfit* was used to compute the solutions of the minimization problem of Section 2.3 corresponding to the three right-hand sides as given by the columns of the matrix B. Table 3 lists the exact solutions and the results obtained when the first and fourth values in Table 1 are replaced by zero.

Table 3

x_1	x_2	x_3	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$
$-1/12$	0	$-1/12$	$-0.0833\,3333$	$0.17_{10}-8$	$-0.0833\,3333$
0	0	0	$-0.58_{10}-8$	$-1.09_{10}-8$	$-1.11_{10}-8$
1/4	0	1/4	$0.2500\,0002$	$1.55_{10}-8$	$0.2500\,0003$
$-1/12$	0	$-1/12$	$-0.0833\,3332$	$0.74_{10}-8$	$-0.0833\,3332$
1/12	0	1/12	$0.0833\,3334$	$0.33_{10}-8$	$0.0833\,3334$
	Residual				
0	$8\sqrt{5}$	$8\sqrt{5}$			

A second example is the 20×21 matrix with entries

$$a_{i,j}=\begin{cases}0 & \text{if } i>j\\ 21-i & \text{if } i=j\\ -1 & \text{if } i<j\end{cases}\qquad \begin{matrix}1\leqq i\leqq 20\\ 1\leqq j\leqq 21\end{matrix}$$

which has orthogonal rows and singular values $\sigma_{21-k}=\sqrt{k(k+1)}$, $k=0,\ldots,20$. Theoretically, the Householder reduction should produce a matrix $J^{(0)}$ with diagonal $-20, 0, \ldots, 0$ and super-diagonal $-\sqrt{20}, \sigma_2, \ldots, \sigma_{20}$. Under the influence of rounding errors a totally different matrix results. However, within working accuracy its singular values agree with those of the original matrix. Convergence is reached after 32 QR transformations and the $\tilde{\sigma}_k$, $k=1,\ldots,20$ are correct within several units in the last digit, $\tilde{\sigma}_{21}=1.61_{10}-11$.

A third example is obtained if the diagonal of the foregoing example is changed to

$$a_{i,i}=1,\qquad 1\leqq i\leqq 20.$$

This matrix has a cluster of singular values, σ_{10} to σ_{19} lying between 1.5 and 1.6, $\sigma_{20}=\sqrt{2}$, $\sigma_{21}=0$. Clusters, in general, have a tendency to reduce the number of required iterations; in this example, 26 iterations were necessary for convergence. $\tilde{\sigma}_{21}=1.49_{10}-8$ is found in eighteenth position and the corresponding column of $\tilde{V}$ differs from the unique solution of the homogeneous system by less than $3.4_{10}-8$ in any component.

A second test was made by Dr. Peter Businger on the CDC 6600.

A third test was performed on the IBM 360/67 at Stanford University. The example used was the 30×30 matrix with entries

$$a_{ij}=\begin{cases}0 & \text{if } i>j\\ 1 & \text{if } i=j\\ -1 & \text{if } i<j.\end{cases}$$

The computed singular values are given in Table 4.

Table 4. *Singular values*

18.2029 0555 7529 2200	6.2231 9652 2604 2340	3.9134 8020 3335 6160	2.9767 9450 2557 7960
2.4904 5062 9660 3570	2.2032 0757 4479 9280	2.0191 8365 4054 5860	1.8943 4154 7685 6890
1.8059 1912 6612 3070	1.7411 3576 7747 9500	1.6923 5654 4395 2610	1.6547 9302 7369 3370
1.6253 2089 2877 9290	1.6018 3335 6666 2670	1.5828 6958 8713 6990	1.5673 9214 4480 0070
1.5546 4889 0109 3720	1.5440 8471 4076 0510	1.5352 8356 5544 9020	1.5279 2951 2160 3040
1.5217 8003 9063 4950	1.5166 4741 2836 7840	1.5123 8547 3899 6950	1.5088 8015 6801 8850
1.5060 4262 0723 9700	1.5038 0424 3812 6520	1.5021 1297 6754 0060	1.5009 3071 1977 0610
1.5002 3143 4775 4370	0.0000 0000 2793 9677		

Note that $\sigma_{30}/\sigma_1 \approx 1.53 \times 10^{-10}$ so that this matrix is very close to being a matrix of rank 29 even though the determinant equals 1.

Acknowledgement. The authors wish to thank Dr. Peter Businger of Bell Telephone Laboratories for his stimulating comments.

References

1. Businger, P., Golub, G.: Linear least squares solutions by Householder transformations. Numer. Math. **7**, 269—276 (1965).
2. Forsythe, G.E., Henrici, P.: The cyclic Jacobi method for computing the principal values of a complex matrix. Proc. Amer. Math. Soc. **94**, 1—23 (1960).
3. — Golub, G.: On the stationary values of a second-degree polynomial on the unit sphere. J. Soc. Indust. Appl. Math. **13**, 1050—1068 (1965).
4. — Moler, C.B.: Computer solution of linear algebraic systems. Englewood Cliffs, New Jersey: Prentice-Hall 1967.
5. Francis, J.: The $Q\,R$ transformation. A unitary analogue to the $L\,R$ transformation. Comput. J. **4**, 265—271 (1961, 1962).
6. Golub, G., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. J. SIAM. Numer. Anal., Ser. B **2**, 205—224 (1965).
7. — Least squares, singular values, and matrix approximations. Aplikace Matematiky **13**, 44—51 (1968).
8. Hestenes, M.R.: Inversion of matrices by biorthogonalization and related results. J. Soc. Indust. Appl. Math. **6**, 51—90 (1958).
9. Kogbetliantz, E.G.: Solution of linear equations by diagonalization of coefficients matrix. Quart. Appl. Math. **13**, 123—132 (1955).
10. Kublanovskaja, V.N.: Some algorithms for the solution of the complete problem of eigenvalues. V. Vyčisl. Mat. i. Mat. Fiz. **1**, 555—570 (1961).
11. Martin, R.S., Reinsch, C., Wilkinson, J.H.: Householder's tridiagonalization of a symmetric matrix. Numer. Math. **11**, 181—195 (1968).
12. Wilkinson, J.: Error analysis of transformations based on the use of matrices of the form $I-2w\,w^H$. Error in digital computation, vol. II, L.B. Rall, ed., p. 77—101. New York: John Wiley & Sons, Inc. 1965.
13. — Global convergence of $Q\,R$ algorithm. Proceedings of IFIP Congress, 1968.

Prof. G.H. Golub
Computer Science Dept.
Stanford University
Stanford, California 94305
USA

Dr. C. Reinsch
Math. Institut
der Techn. Hochschule
8000 München 2
Arcisstr. 21

Numer. Math. 14, 421—434 (1970)

Numerical Integration of Products of Fourier and Ordinary Polynomials*

D. G. Bettis

Received June 6, 1969

Abstract. Sets of coefficients for four finite difference methods of numerical integration are presented that will integrate without truncation error products of fourier and ordinary polynomials. These sets are formulated such that they are free from computational difficulties.

I. Introduction

In a previous paper [4] sets of modified integration coefficients for the Cowell method of numerical integration of order six were given which had the property that they would integrate without truncation error products of a fourier polynomial and an ordinary polynomial. These sets of coefficients were characterized by their explicit formulation. The purpose of this paper is the extension of the set of modified Cowell coefficients to any order of the integration method, as well as the development of similar sets of modified integration coefficients for the Störmer, Adams-Moulton, and the Adams-Bashforth methods of numerical integration. These modified coefficients will be given explicitly in a form such that they can be computed from a simple algorithm which does not suffer from numerical difficulties.

The four methods of numerical integration based upon backward differences will be referred to as follows [2]: 1) Cowell — the implicit method for differential equations of the second order in which the first derivatives are absent; 2) Störmer — the explicit method for these second order differential equations; 3) Adams-Moulton — the implicit method for differential equations of the first order; and 4) Adams-Bashforth — the explicit method for differential equations of the first order.

II. Integration Formulae

For the integration of a function $f(t)$, tabulated at equally spaced values of the independent variable t with the step-length h, any ascending diagonal of a difference table can be used for the development of a method of numerical integration [4]. In such a table, let m be an integer, $t_m = mh$, $f_m = f(t_m)$; and let $\Delta^0 f(m) = f_m$, $\Delta f(m-\frac{1}{2}) = f_m - f_{m-1}$, $\Delta^2 f(m) = f_{m+1} - 2f_m + f_{m-1}$, ect. For a differential equation of the form

$$\frac{d^2 x(t)}{dt} = f(x, t)$$

* Part of this work was formulated earlier by the author in a dissertation presented to Yale University in partial fulfillment for the degree of Doctor of Philosophy.

the double integration may be obtained by

$$x_1 - 2x_0 + x_{-1} = h^2 \sum_{k=0}^{\infty} \alpha_k \Delta^k f\left(p - \frac{k}{2}\right) \tag{1.1}$$

where $\alpha_0 = 1$, $p = 1$ corresponds to the classical Cowell formula, and $p = 0$ represents the Störmer formula.

For the integration of a differential equation of the first order,

$$\frac{dx(t)}{dt} = f(x, t)$$

a suitable integration formula is

$$x_1 - x_0 = h \sum_{k=0}^{\infty} \alpha_k \Delta^k f\left(p - \frac{k}{2}\right). \tag{1.2}$$

When $p = 1$ the formula is the classical Adams-Moulton method, and when $p = 0$ it is the classical Adams-Bashforth method. Since the sets of α's are assumed to be constant in these classical methods, the solution of the differential equations is expressed as a linear combination of the differences of a particular diagonal of the difference table. The coefficients α are different for the four methods of integration. However, with this distinction well in mind, no ambiguity will result if the same symbol is used for all four of the integration methods.

In order to determine the four sets of coefficients α, assume that

$$f(t) = Z^{t/h} \tag{2}$$

where Z represents any fixed complex number. Then it follows that for the double integration

$$\int_0^h \int_{\tau-h}^{\tau} Z^{\tau/h} \, d\tau \, dt = Z \left[\frac{h(1 - Z^{-1})}{\log Z}\right]^2,$$

and for the single integration

$$\int_0^h Z^{t/h} = \frac{h}{\log Z} (1 - Z^{-1}) Z.$$

Letting

$$\zeta = 1 - Z^{-1}$$

the integration formulae (1) become

$$\left[\frac{\zeta}{\log(1-\zeta)}\right]^2 = P(\zeta), \tag{3.1}$$

$$\frac{1}{1-\zeta}\left[\frac{\zeta}{\log(1-\zeta)}\right]^2 = P(\zeta), \tag{3.2}$$

$$\frac{\zeta}{\log(1-\zeta)} = P(\zeta), \tag{3.3}$$

$$\frac{\zeta}{(1-\zeta)\log(1-\zeta)} = P(\zeta), \tag{3.4}$$

where $P(\zeta)$ is defined as

$$P(\zeta) = \sum_{k=0}^{\infty} \alpha_k \zeta^k, \tag{4}$$

and where the integers 1, 2, 3, 4 denote the Cowell, Störmer, Adams-Moulton, and the Adams-Bashforth methods, respectively.

Upon expanding the left-hand sides of (3) in powers of ζ, with $|\zeta| < 1$, the coefficients of similar terms of ζ may be compared, yielding, for the first six coefficients the following sets [2]:

Cowell method —

$$\begin{aligned} &\alpha_1 = -1, && \alpha_2 = \frac{1}{12}, && \alpha_3 = 0, \\ &\alpha_4 = -\frac{1}{240}, && \alpha_5 = -\frac{1}{240}, && \alpha_6 = \frac{-221}{60\,480}; \end{aligned} \tag{5.1}$$

Störmer method —

$$\begin{aligned} &\alpha_1 = 0, && \alpha_2 = \frac{1}{12}, && \alpha_3 = \frac{1}{12}, \\ &\alpha_4 = \frac{19}{240}, && \alpha_5 = \frac{3}{40}, && \alpha_6 = \frac{863}{12\,096}; \end{aligned} \tag{5.2}$$

Adams-Moulton method —

$$\begin{aligned} &\alpha_1 = -\frac{1}{2}, && \alpha_2 = -\frac{1}{12}, && \alpha_3 = -\frac{1}{24}, \\ &\alpha_4 = \frac{-19}{720}, && \alpha_5 = \frac{-27}{1440}, && \alpha_6 = \frac{-863}{60\,480}; \end{aligned} \tag{5.3}$$

Adams-Bashforth method —

$$\begin{aligned} &\alpha_1 = \frac{1}{2}, && \alpha_2 = \frac{5}{12}, && \alpha_3 = \frac{9}{24}, \\ &\alpha_4 = \frac{251}{720}, && \alpha_5 = \frac{475}{1440}, && \alpha_6 = \frac{19\,087}{60\,480}. \end{aligned} \tag{5.4}$$

The order, n, of the integration method is defined by the subscript of the highest coefficient α_n retained in (1). It is important to remember that the classical sets of integration coefficients are based upon the expansion of the left-hand sides of (3) in an infinite power series in terms of the variable ζ. If, in application, (1) are used with the order chosen such that either the highest coefficients are zero, or the difference Δ^n of the function being integrated vanishes identically, then (1) will integrate the function exactly. However, in practice these two conditions are usually not satisfied. In effect, the integration of the function is being approximated by the truncated polynomial of the right-hand side of (3),

$$P_n(\zeta) = \sum_{k=0}^{n} \alpha_k \zeta^k. \tag{6}$$

Therefore, the integration of a function by the use of (1) with a finite set of coefficients will result in an error because the formulae are based upon an approximating polynomial instead of an exact polynomial.

While formulating a method to reduce these truncation errors for the four classical methods of numerical integration when they are applied to functions which are products of fourier polynomials and ordinary polynomials, the basic formulae (1) will not be altered, but the coefficients α will be modified.

III. Formulation of the Problem

Before deriving the modified coefficients which will integrate without truncation error the products of fourier and ordinary polynomials, the modified coefficients will be obtained which will integrate only fourier polynomials exactly. The final sets of coefficients which will integrate the products of the fourier and the ordinary polynomials will be obtained as limiting cases of these first sets of coefficients.

Much insight into the general problem of deriving the modified coefficients results from the determination of the coefficients which will integrate the functions $\sin \omega t$ and $\cos \omega t$. Thus, for (2) let $f(t) = \exp(i \omega t)$, and introducing the notation $2\sigma = \omega h$, Z in (2) becomes

$$Z = \exp(2i\sigma),$$

and ζ becomes

$$\zeta = 1 - \exp(-2i\sigma).$$

Therefore, Eqs. (3) reduce to:

$$L_i(\sigma) = P_n(\zeta), \quad i = 1, 2, 3, 4, \tag{7}$$

where

$$L_1(\sigma) = \left(\frac{\sin \sigma}{\sigma}\right)^2 \exp(-2i\sigma),$$

$$L_2(\sigma) = \left(\frac{\sin \sigma}{\sigma}\right)^2,$$

$$L_3(\sigma) = \frac{\sin \sigma}{\sigma} \exp(-i\sigma),$$

$$L_4(\sigma) = \frac{\sin \sigma}{\sigma} \exp(i\sigma).$$

The coefficients α must be chosen such that (7) are satisfied for the given frequency ω, as well as for $-\omega$, if the four integration methods of order n are to integrate the functions $f(t) = \sin \omega t$ and $f(t) = \cos \omega t$ exactly.

For the integration methods of order two the coefficients α_1 and α_2 must be chosen such that the $L_i(\sigma)$ are identical to

$$1 + \alpha_1 \zeta + \alpha_2 \zeta^2,$$

or

$$1 + \alpha_1 [1 - \exp(-2i\sigma)] + \alpha_2 [1 - \exp(-2i\sigma)]^2.$$

For the methods of order two, the modified coefficients can be determined by solving the two equations which result from (7) when the two frequencies ω and $-\omega$ are considered. If σ is not zero, or if, for the two Adams methods, σ is not an odd integer multiple of $\pi/2$, then the solution of the two equations yields for the

Cowell method —

$$\alpha_1 = -1, \qquad \alpha_2 = \lambda(u); \tag{8.1}$$

Störmer method —

$$\alpha_1 = \lambda^*(u) - 1, \qquad \alpha_2 = \lambda(u); \tag{8.2}$$

Adams-Moulton method —

$$\alpha_1 = \mu^*(u) - 1, \qquad \alpha_2 = \mu(u); \tag{8.3}$$

Adams-Bashforth method —

$$\alpha_1 = (3-u)\,\mu^* - 1, \qquad \alpha_2 = \mu(u) + \mu^*(u); \tag{8.4}$$

where

$$\sigma = \frac{\omega h}{2}, \qquad u = 4\sin^2\sigma,$$

and where

$$\lambda^*(u) = \left(\frac{\sin\sigma}{\sigma}\right)^2, \qquad \mu^*(u) = \frac{\sin\sigma}{2\sigma\cos\sigma},$$

$$\lambda(u) = \frac{1}{4}\left(\frac{1}{\sin^2\sigma} - \frac{1}{\sigma^2}\right),$$

$$\mu(u) = \frac{1}{4}\left(\frac{1}{\sin^2\sigma} - \frac{1}{\sigma\sin\sigma\cos\sigma}\right).$$

With these choices of α_1 and α_2 the four integration methods of second order will integrate the functions $\sin\omega t$ and $\cos\omega t$ exactly, except for any round-off error which accumulates in the algorithm.

Henceforth, the coefficients which have been adjusted in the above manner, will be referred to as modified coefficients, and the coefficients of (5) will be referred to as the classical coefficients. For the modified coefficients two subscripts, i and n, will denote the i-th coefficient of the set and the order of the method, respectively. As σ becomes very small, the above coefficients approach the corresponding classical coefficients.

If the order of the integration method is greater than two, then the $\alpha_{i,n}$ $(i = 1, 2, \ldots, n)$ coefficients can be determined such that the functions $\cos\omega t$ and $\sin\omega t$ are integrated exactly. It will be assumed that n is even and equal to 2ν. Thus, there will exist ν distinct frequencies $\omega_1, \omega_2, \ldots, \omega_\nu$ such that the 2ν coefficients will integrate the functions $\cos\omega_k t$ and $\sin\omega_k t$ $(k = 1, 2, \ldots, \nu)$ exactly, if (7) are satisfied for the corresponding σ_k and for $-\sigma_k$. Since the right-hand sides of (7) are polynomials in the variable ζ, the above condition is equivalent to the following problem of polynomial interpolation:

Given $2\nu+1$ points ζ_k, $k = -\nu, \ldots, -1, 0, +1, \ldots, +\nu$, where $\zeta_k = 1 - \exp(-2i\sigma_k)$, and where the points are on a circle which has a unit radius in the complex ζ-plane, construct a polynomial $P_n(\zeta)$, $n = 2\nu$, such that $L_i(\sigma_k) = P_n(\zeta)$, $i = 1, 2, 3, 4$, and where $\sigma_{-k} = -\sigma_k$.

Once the coefficients α are determined such that the above conditions are satisfied, the integration methods will integrate without truncation error not

only the functions $\sin \omega_k t$ and $\cos \omega_k t$, but also the linear combination

$$a_0 + \sum_{k=1}^{\nu} (a_k \cos \omega_k t + b_k \sin \omega_k t). \tag{9}$$

The constant a_0 is integrated exactly because the polynomial $P_n(\zeta)$ begins with unity.

IV. A Special Case

When only two of the n coefficients are modified, the remainder of the coefficients retaining their classical values, the resulting set of coefficients will integrate exactly the function

$$Q_{n-2}(t) + a \cos \omega t + b \sin \omega t, \tag{10}$$

where $Q_m(t)$ is an ordinary polynomial of degree m in the variable t. In this special case there are given two points $\zeta_k = 1 - \exp(-2i\sigma_k)$, $k = -1, +1$, on a unit circle in the complex ζ-plane passing through the orgin, where the remaining $n-2$ points are located. The problem is the construction of a polynomial $P_n(\zeta)$ such that the relations (7) are satisfied for the two values of k. This problem can be solved by the technique which was used for the modification of the coefficients for the integration methods of order two. Thus, the problem reduces to the solution of the two equations ($k = -1, +1$):

$$L_i(\sigma_k) = P_n(\zeta_k), \quad i = 1, 2, 3, 4, \tag{11}$$

for two of the coefficients, for example $\alpha_{n,n}$ and $\alpha_{n-1,n}$.

Before solving for the two unknown coefficients it is advantageous to introduce two sequences of polynomials $R_m(u)$ and $S_m(u)$ defined by the recurrence relations [4]:

$$\begin{aligned} R_{m+1} &= u(R_m - R_{m-1}), \quad R_0 = 2, \quad R_1 = u, \\ S_{m+1} &= u(S_m - S_{m-1}), \quad S_0 = 0, \quad S_1 = u. \end{aligned} \tag{12}$$

For $m = 0, 1, 2, \ldots$, the following relations can be proved by induction:

$$\begin{aligned} \sin 2m\sigma &= \frac{(-1)^{m-1} \sin 2\sigma \, S_{2m}}{u^{m+1}}, \\ \sin (2m+1)\sigma &= \frac{(-1)^m \sin \sigma \, R_{2m+1}}{u^{m+1}}, \end{aligned} \tag{13}$$

and

$$\begin{aligned} S_{2m} &= (-1)^{m-1} u^{m+1} (m + Q^*_{m-1}), \\ R_{2m+1} &= (-1)^m u^{m+1} (2m + 1 + Q^*_m), \end{aligned} \tag{14}$$

where $Q^*_p(u)$ is a polynomial with constant coefficients of degree p in the variable u with the term of degree zero absent. Also the following preliminary relations will be needed:

$$\begin{aligned} \zeta_k^m - \zeta_{-k}^m &= \frac{\tau \, S_m(u_k)}{u_k}, \\ \zeta_k^m + \zeta_{-k}^m &= R_m(u_k), \end{aligned} \tag{15}$$

where $\tau = 2i \sin 2\sigma$.

Solving (11) for $\alpha_{n,n}$ and $\alpha_{n-1,n}$ gives

$$\alpha_{p,n} = \frac{1}{u^{n-1}}\left[\frac{L_{i,1}\zeta_{-1}^{q} - L_{i,-1}\zeta_{1}^{q}}{\tau} + \sum_{k=0}^{n-2} \alpha_{k,n}\, u^{k-1}\left(\frac{\zeta_{1}^{q-k} - \zeta_{-1}^{q-k}}{\tau}\right)\right],$$

where $\alpha_{0,n} = 1$, $q = n-1$ when $p = n$, $q = n$ when $p = n-1$, and $L_{i,k} = L_i(\sigma_k)$.

Using (15), the definition of ζ_k, and the relation,

$$\exp(i\sigma_k) = (\zeta_k)^{-1} 2i \sin\sigma_k,$$

and defining the first term in the bracket as $F^*_{p,i}(u)$, then

$$F^*_{p,1} = \lambda^*(u)\, S_{q-2},$$

$$F^*_{p,2} = -\lambda^*(u)\, S_q/u,$$

$$F^*_{p,3} = -\mu^*(u)\, R_{q-1},$$

$$F^*_{p,4} = \mu^*(u)\, R_{q+1}/u.$$

Therefore, if the classical coefficients are used for $\alpha_{l,n}$, $l = 1, 2, \ldots, n-2$, then the last two coefficients are

$$\alpha_{p,n} = \frac{(-1)^{n-p}}{u^{n-1}}\left[F^*_{p,i}(u) + \sum_{k=0}^{n-2} \alpha_{k,n}\, u^{k-1} S_{q-k}\right], \tag{16}$$

$$p = n,\ q = n-1;\quad p = n-1,\ q = n.$$

With these coefficients the integration methods will integrate the function (10) exactly.

V. Modified Coefficients for Distinct Frequencies

To integrate the function (9) without any truncation error requires the solution of the problem of interpolation given in Section III. Any of the classical interpolation formulae are available for this purpose. Here the technique will be to obtain the two highest coefficients by an application of the Lagrangian interpolation method:

$$L_i(\sigma) = \sum_{k=-\nu}^{\nu} L_{i,k} \prod_{\substack{m=-\nu \\ m\neq k}}^{\nu} \frac{\zeta - \zeta_m}{\zeta_k - \zeta_m}, \qquad n = 2\nu. \tag{17}$$

The lower coefficients could be obtained by the same technique, but this results in a rather tedious procedure that involves computation in the complex variable ζ. Therefore, a more efficient algorithm in the real field is developed in which the lower coefficients are determined by recurrence with respect to the coefficients of lower order. The final goal is the determination of the two highest coefficients in explicit form for any order. Thus, after the remaining coefficients are found by the recurrence method, the complete set of modified coefficients for the four methods will be readily available.

The derivation of the expression for the highest coefficient for the four methods of any given even order will be given. The modified coefficient $\alpha_{n,n}$

equals the coefficient of the ζ^n term of the Lagrangian interpolation formula:

$$L_{i,0}\prod_{k=1}^{\nu}\frac{1}{u_k}+\sum_{k=1}^{\nu}\frac{L_{i,k}\,\zeta_{-k}(\zeta_{-k}-1)^{\nu-1}-L_{i,-k}\,\zeta_k(\zeta_k-1)^{\nu-1}}{(\zeta_k-\zeta_{-k})\,u_k}\Pi^*, \tag{18}$$

where

$$\Pi^*=\prod_{\substack{m=1\\ m\neq k}}^{\nu}\frac{1}{u_k-u_m}.$$

The expression (18) can be expressed in terms of u, $\lambda(u)$, and $\mu(u)$ for the four integration methods.

Since, for the Cowell method,

$$L_{1,k}=\frac{u_k}{4\sigma_k^2}(1-\zeta_k),$$

the right-hand side of (18) becomes:

$$\sum_{k=1}^{\nu}\frac{(1-\zeta_k)\,\zeta_{-k}(\zeta_{-k}-1)^{\nu-1}-(1-\zeta_{-k})\,\zeta_k(\zeta_k-1)^{\nu-1}}{4\sigma_k^2(\zeta_k-\zeta_{-k})}\Pi^*. \tag{19}$$

Since

$$1-\zeta_k=-(\zeta_{-k}-1)^{-1},$$

the numerator of (19) becomes

$$-\zeta_{-k}(\zeta_{-k}-1)^{\nu-2}+\zeta_k(\zeta_k-1)^{\nu-2}$$

or, because $\zeta_k=1-\exp(2i\,\sigma_k)$, the numerator reduces to

$$(-1)^{\nu-1}[2i\sin 2(\nu-2)\,\sigma_k]+(-1)^{\nu}[2i\sin 2(\nu-1)\,\sigma_k].$$

Using the definitions (13), the above expression simplifies to

$$\tau\left[\frac{S_{2(\nu-2)}}{u^{\nu-1}}+\frac{S_{2(\nu-1)}}{u^{\nu}}\right],$$

and, since the denominator is equal to $4\sigma_k^2\tau$, (18) becomes

$$\sum_{k=1}^{\nu}\left[\frac{S_{2(\nu-2)}}{u_k^{\nu-1}}+\frac{S_{2(\nu-1)}}{u_k^{\nu}}\right]\frac{\Pi^*}{4\sigma_k^2}. \tag{20}$$

With the aid of the first formula of (14), the expression in brackets can be shown to be of the form

$$(-1)^{\nu}[1+Q^*_{\nu-2}(u_k)],$$

where

$$Q^*_m(u)=q_1u+q_2u^2+\cdots+q_mu^m.$$

Therefore, by using the finite expansion of $1/(u_1u_2\ldots u_\nu)$ from Section VII, (19) becomes of the form

$$(-1)^{\nu-1}\sum_{k=1}^{\nu}\left[\frac{1+c_1u_k+\cdots+c_{\nu-2}u_k^{\nu-2}}{u_k}-\frac{1+q_1u_k+\cdots+q_{\nu-2}u_k^{\nu-2}}{4\sigma_k^2}\right]\Pi^*. \tag{21}$$

Since the constants c are arbitrary, they can be chosen such that the numerators of the two terms in the brackets of (21) are identical, i.e., the coefficients of the polynomial

$$1 + c_1 u_k + \cdots + c_{\nu-2} u_k^{\nu-2}$$

are selected so that the polynomial equals

$$\frac{S_{2(\nu-2)}}{u_k^{\nu-1}} + \frac{S_{2(\nu-1)}}{u_k^{\nu}} = \frac{S_{2\nu-3}}{u_k^{\nu-1}}.$$

Thus, (21) can be expressed as

$$-\sum_{k=1}^{\nu} \frac{S_{2\nu-3}(u_k)}{u_k^{\nu-1}} \lambda(u_k) \Pi^*,$$

which is the $\alpha_{n,n}$ coefficient for the case of distinct frequencies for the Cowell method.

Defining $F_{n,1}(u)$ as

$$F_{n,1}(u) = -\frac{S_{2\nu-3}(u)}{u^{\nu-1}} \lambda(u), \tag{22.1}$$

the $\alpha_{n,n}$ coefficient for the Cowell method becomes the $(\nu-1)$-th divided difference of the function $F_{n,1}(u)$ at the nodes $u_1, u_2, \ldots, u_\nu$. In a similar manner it follows that the $\alpha_{n,n}$ coefficient for the Störmer $(i=2)$ and the Adams-Moulton $(i=3)$ methods is the $(\nu-1)$-th divided difference of the function $F_{n,i}(u)$, where

$$F_{n,2}(u) = \frac{S_{2\nu-1}(u)}{u^{\nu}} \lambda(u), \tag{22.2}$$

$$F_{n,3}(u) = \frac{R_{2\nu-2}(u)}{2u^{\nu-1}} \mu(u). \tag{22.3}$$

Likewise, for the Adams-Bashforth method $(\alpha_{n,n} - \frac{1}{2})$ is the $(\nu-1)$-th divided difference of $F_{n,4}(u)$, where

$$F_{n,4}(u) = -\frac{R_{2\nu}(u)}{2u^{\nu}} \mu(u). \tag{22.4}$$

The factor one-half appears in this method because a term of degree $(\nu-1)$ is necessary in the finite expansion of the first term of (18).

The second highest coefficient equals the coefficient of the ζ^{n-1} term of the Lagrangian interpolation formula:

$$\Sigma^* - L_{i,0} \prod_{k=1}^{\nu} \frac{1}{u_k} + \sum_{k=1}^{\nu} \frac{L_{i,k}\zeta_{-k}(\zeta_{-k}-1)^{\nu-2} - L_{i,-k}\zeta_k(\zeta_k-1)^{\nu-2}}{(\zeta_k - \zeta_{-k})u_k} \Pi^*, \tag{23}$$

$$i = 1, 2, 3, 4,$$

where

$$\Sigma^* = \left(1 - \sum_{k=1}^{\nu} u_k\right) \alpha_{n,n}.$$

With the aid of (12), (13), and (14), and with a development analogous to that for the highest coefficient, $(\alpha_{n-1,n} - \Sigma^*)$ becomes the $(\nu-1)$-th divided difference

of the function $F_{n-1,i}(u)$, $i=1, 2, 3, 4$, where

$$
\begin{aligned}
F_{n-1,1}(u) &= -\frac{S_{2\nu-5}(u)}{u^{\nu-2}}\lambda(u), \\
F_{n-1,2}(u) &= \frac{S_{2\nu-3}(u)}{u^{\nu-1}}\lambda(u), \\
F_{n-1,3}(u) &= \frac{R_{2\nu-4}(u)}{2u^{\nu-2}}\mu(u), \\
F_{n-1,4}(u) &= -\frac{R_{2\nu-2}(u)}{2u^{\nu-1}}\mu(u).
\end{aligned} \tag{24}
$$

In order to obtain the remaining coefficients $\alpha_{m,n}$, $m=1, 2, \ldots, n-2$, assume that the set of coefficients of order $n-2$ are known and that the two highest coefficients have been computed. Consider two polynomials $P_n(\zeta)$ and $P_{n-2}(\zeta)$ interpolated at the nodes

$$\zeta_0, \zeta_{\pm 1}, \ldots, \zeta_{\pm\nu}, \tag{25}$$

and

$$\zeta_0, \zeta_{\pm 1}, \ldots, \zeta_{\pm l}, \quad l=\nu-1, \tag{26}$$

respectively. The polynomial having the points (26) as zeros is

$$\zeta(\zeta^2-u_1\zeta+u_1)(\zeta^2-u_2\zeta+u_2)\ldots(\zeta^2-u_l\zeta+u_l).$$

This polynomial which vanishes at the $n-2$ zeros, multiplied by a linear factor, is the difference of the two polynomials $P_n(\zeta)$ and $P_{n-2}(\zeta)$,

$$P_n(\zeta)-P_{n-2}(\zeta)=\zeta(\zeta^2-u_1\zeta+u_1)(\zeta^2-u_2\zeta+u_2)\ldots(\zeta^2-u_l\zeta+u_l)(r\zeta-s). \tag{27}$$

The parameters r and s can be obtained as a function of the two highest coefficients by first expressing the left-hand sides of (27) in terms of the modified coefficients of order n and $n-2$, respectively, and by then comparing the coefficients of ζ^n and ζ^{n-1}, yielding

$$r=\alpha_{n,n}, \qquad s=\alpha_{n-1,n}+\alpha_{n,n}\sum_{k=1}^{l}u_k.$$

With these values for r and s, (27) becomes

$$P_n(\zeta)=P_{n-2}(\zeta)+\left[\alpha_{n,n}\zeta+\alpha_{n-1,n}+\alpha_{n,n}\sum_{k=1}^{l}u_k\right]\prod_{k=1}^{l}(\zeta^3-u_k\zeta^2+u_k\zeta). \tag{28}$$

The lower coefficients can now be obtained by comparing the coefficients of similar terms of ζ in (28). This technique is easily adaptable to a computer subroutine.

VI. Modified Coefficients for Confluent Frequencies

The modified coefficients previously developed will integrate without truncation error the function (9). However, if two or more of the frequencies ω_k approach a common limit, for example if $\omega_1\to\omega_2\to\cdots\to\omega_m$, the modified coefficients

for this case of confluent frequencies will integrate exactly the function

$$a_0 + Q_{m-1}(t)\left[\sum_{k=m+1}^{\nu}(a_k\cos\omega_k t + b_k\sin\omega_k t)\right], \tag{29}$$

i.e., the product of an ordinary polynomial and a fourier polynomial.

In order to obtain the sets of modified coefficients which will integrate the above function, the limits must be obtained of the modified coefficients for distinct frequencies as the m frequencies approach the value of confluency. Since the modified coefficients for distinct frequencies are the divided differences of the function $F_{p,i}(u)$, $p = n, n-1$, and $i = 1, 2, 3, 4$, at the nodes $u_1, u_2, \ldots, u_\nu$, the desired limits are available from the well established theory of divided differences [3].

Sets of these modified coefficients will be given without proof for three of the more important limiting cases: a) two, b) $\nu - 1$, and c) all of the ν frequencies approach a common limit. The first case is applicable when the two highest coefficients for distinct frequencies suffer from a loss of significant digits because the differences of u_1 and u_2 in their denominators become small. As the confluent frequencies approach zero in the second case, the coefficients approach those of the special case of Section IV. The third case is characterized by the lack of the troublesome factors of u in the denominators of the two highest coefficients. Only the two highest coefficients will be given since the lower coefficients may be computed by the recurrence technique of Section V.

In the first case assume that ω_1 and ω_2 approach the common limit ω. If $G_n(u)$ is defined as

$$G_n(u) = \prod_{k=3}^{\nu}(u - u_k), \qquad n \geqq 6,$$

then

$$\alpha_{p,n} = K_{p,i} + D\left[\frac{F_{p,i}(u)}{G_n(u)}\right] + \sum_{k=3}^{\nu} F_{p,i}(u_k)\,\Pi^*, \qquad p = n, n-1, \tag{30}$$

with the definitions

$$K_{n,i} = 0 \quad \text{if} \quad i = 1, 2, 3, \qquad K_{n,4} = \tfrac{1}{2},$$

and

$$K_{n-1,i} = \Sigma^* \quad \text{for} \quad i = 1, 2, 3, 4,$$

and where $D = d/du$. For the second case let the first $(\nu - 1)$ frequencies approach the common limit ω, then

$$\alpha_{p,n} = K_{p,i} + \frac{(-1)^\nu}{(u-u_\nu)^{\nu-1}}\sum_{k=2}^{\nu-2}\frac{(u_\nu - u)^k D^k F_{p,i}(u)}{k!} + \frac{F_{p,i}(u_\nu)}{(u_\nu - u)^{\nu-1}}, \qquad p = n, n-1, \tag{31}$$

where the operator D is defined as

$$D^0 = 1; \qquad D^k = \frac{d^k}{du^k}, \qquad k = 1, 2, \ldots.$$

For the third case all the frequencies approach the limit ω. Here the highest coefficients become

$$\alpha_{p,n} = K_{p,i} + \frac{D^{\nu-1}F_{p,i}(u)}{(\nu-1)!}, \qquad p = n, n-1. \tag{32}$$

VII. A Finite Expansion

The development of the two highest coefficients for the case of distinct frequencies requires an expansion of the function $1/(u_1 u_2 \dots u_\nu)$.

The $(\nu-1)$-th divided difference of the function $g(u)=1/u$ is [3]:

$$[u_1 u_2 \dots u_\nu] = \frac{(-1)^{\nu-1}}{\prod_{k=1}^{\nu} u_k}, \tag{33}$$

where the $(\nu-1)$-th divided difference of a function $g(u)$ is defined by

$$[u_1 u_2 \dots u_\nu] = \sum_{k=1}^{\nu} g(u_k) \Pi^*, \quad \Pi^* = \prod_{\substack{m=1 \\ m \neq k}}^{\nu} \frac{1}{(u_k - u_m)}. \tag{34}$$

Combining (33) and (34) gives

$$\frac{1}{\prod_{k=1}^{\nu} u_k} = (-1)^{\nu-1} \sum_{k=1}^{\nu} \frac{1}{u_k} \Pi^*. \tag{35}$$

Since the $(\nu-1)$-th divided difference of a polynomial of degree $(\nu-2)$ is zero, then

$$\sum_{k=1}^{\nu} \frac{c_1 u_k + c_2 u_k^2 + \dots + c_{\nu-1} u_k^{\nu-1}}{u_k} \Pi^* = 0. \tag{36}$$

Furthermore, the $(\nu-1)$-th divided difference of a polynomial of degree $\nu-1$ equals the coefficient of the term of the polynomial of degree $\nu-1$, or

$$\sum_{k=1}^{\nu} \frac{c_\nu u_k^\nu}{u_k} \Pi^* = c_\nu. \tag{37}$$

Adding (35), (36), and (37) yields the desired finite expansion of the function $1/(u_1 u_2 \dots u_\nu)$:

$$\frac{1}{\prod_{k=1}^{\nu} u_k} = -c_\nu + \sum_{k=1}^{\nu} \frac{(-1)^{\nu-1}(1 + c_1 u_k + \dots + c_{\nu-1} u_k^{\nu-1}) + c_\nu u_k^\nu}{u_k} \Pi^*, \tag{38}$$

where the constants c are arbitrary.

VIII. Some Power Expansions

During computation the functions $\lambda^*(u)$, $\lambda(u)$, $\mu^*(u)$, and $\mu(u)$, and their derivatives will suffer from a loss of significant digits if the variable σ becomes small. This loss of precision may be avoided by expanding the functions in a power series in terms of the variable u.

The expansions of $\lambda^*(u)$ and $\lambda(u)$ are presented in [4]:

$$\lambda^*(u) = 1 - \frac{1}{12} u - \frac{1}{240} u^2 - \frac{31}{60\,480} u^3 - \dots,$$

$$\lambda(u) = \frac{1}{12} + \frac{1}{240} u + \frac{31}{60\,480} u^2 + \frac{289}{3\,628\,800} u^3 + \dots.$$

In order to obtain similar expansions for $\mu^*(u)$ and $\mu(u)$ as a function of the variable u consider the following integration formula [1]:

$$(39)\quad \Delta x\left(\frac{1}{2}\right)=\int_0^h f(t)\,dt=h\left[1+\frac{1}{2}\Delta f\left(\frac{1}{2}\right)-\frac{1}{12}\Delta^2 f(0)-\frac{1}{24}\Delta^3 f\left(\frac{1}{2}\right)\right.$$
$$\left.+\frac{11}{120}\Delta^4 f(0)+\frac{11}{1440}\Delta^5 f\left(\frac{1}{2}\right)-\frac{191}{60\,480}\Delta^6 f(0)-\frac{191}{120\,960}\Delta^7 f\left(\frac{1}{2}\right)+\cdots\right].$$

The function to which this formula will be applied is $f(t)=\cos\omega t$. Since $t_m=mh$, and since $2\sigma=\omega h$, it follows that $f_m=\cos 2m\sigma$. Also, since

$$\cos 2(m+1)\sigma-\cos 2m\sigma=-2\sin(2m+1)\sigma\sin\sigma,$$

then

$$\Delta f(m)=-2\sin 2\sigma\sin 2m\sigma-\frac{u}{2}\cos 2m\sigma.$$

Furthermore,

$$\Delta^{2k-1}f(m)=(-1)^k\left[u^{k-1}\sin 2\sigma\sin 2m\sigma+\frac{u^k}{2}\cos 2m\sigma\right],$$

or, the odd first differences of $\cos\omega t$ are

$$\Delta^{2k-1}f(0)=\frac{1}{2}(-u)^k.$$

Likewise, since

$$\cos 2(m+1)\sigma+\cos 2(m-1)\sigma=2\cos 2\sigma\cos 2m\sigma,$$

then

$$\Delta^2 f(m)=-u\cos 2m\sigma,$$

and, in general,

$$\Delta^{2k}f(m)=(-u)^k\cos 2m\sigma.$$

Thus, the even central differences of $\cos\omega t$ are

$$\Delta^{2k}f(0)=(-u)^k.$$

From the integral

$$\int_0^h\cos\omega t\,dt=\frac{1}{\omega}\sin\omega h,$$

it follows that

$$\Delta x\left(\frac{1}{2}\right)=\frac{\sin\omega h}{\omega}=\frac{h\sin\sigma\cos\sigma}{\sigma}.$$

With these result (39) becomes

$$\frac{\sin\sigma\cos\sigma}{\sigma}=1-\frac{1}{6}u-\frac{1}{180}u^2-\frac{1}{1512}u^3-\frac{23}{226\,800}u^4-\cdots.$$

Using this relation the expansions of $\mu^*(u)$ and $\mu(u)$ are:

$$\mu^*(u)=\frac{1}{2}+\frac{1}{24}u+\frac{11}{1440}u^2+\frac{191}{120\,960}u^3+\cdots,$$

$$\mu(u)=-\frac{1}{12}-\frac{11}{720}u-\frac{191}{60\,480}u^2-\frac{2497}{3\,628\,800}u^3-\cdots.$$

IX. Elimination of Numerical Difficulties

The two highest coefficients suffer from a loss of significant digits during their computation when the differences of $u_1, u_2, \ldots, u_m$, $m \geqq 2$, in the denominators become small. This occurs when m of the frequencies approach a common limit. This difficulty may be avoided by using the set of coefficients for $l+m$ confluent frequencies, where l denotes the number of confluent frequencies of the original set of coefficients. For example, if this situation appears because ω_1 and ω_2 approach a common value when the modified coefficients for distinct frequencies are computed, the set of coefficients (30) should be used.

When the variable σ becomes small the functions $\lambda^*(u)$, $\lambda(u)$, $\mu^*(u)$, $\mu(u)$ and their derivatives become indeterminant. If this occurs the power expansions from Section VIII will eliminate the difficulty.

If $\sigma = (2m+1)\frac{\pi}{2}$, $m = 0, 1, 2, \ldots$, the functions $\mu^*(u)$ and $\mu(u)$ and their derivatives, and the derivatives of the functions $\lambda^*(u)$ and $\lambda(u)$ become infinite. For a given problem the value of ω is specified, but the value of h can be varied. Therefore, since $2\sigma = \omega h$, a suitable choice of the step-length will avoid the problem of σ being an odd integer multiple of $\pi/2$.

Acknowledgments. The author gratefully acknowledges the advice and the practical insight which Professor E. Stiefel has provided during the preparation of this paper. The financial support of the Jet Propulsion Laboratory of Pasadena, California, during the author's summers of employment is acknowledged.

References

1. Bashforth, F., Adams, J. C.: Theory of capillary action. Cambridge: University Press 1882.
2. Henrici, P.: Discrete variable methods in ordinary differential equations. New York: John Wiley & Sons, Inc. 1962.
3. Steffensen, J. F.: Interpolation. Baltimore: Williams & Wilkins Co. 1927.
4. Stiefel, E., Bettis, D. G.: Stabilization of cowell's method. Numer. Math. **13**, 154–175 (1969).

Dr. D. G. Bettis
Eidgenössische Technische Hochschule
Institut für Angewandte Mathematik
Schmelzbergstraße 28
8044 Zürich, Schweiz

Numer. Math. 14, 435—447 (1970)

Stability Analysis of a Difference Scheme for the Navier-Stokes Equations

N. G. Campbell

Received June 9, 1969

Abstract. An application of stability analysis by the energy method is made to a practical problem of unsteady viscous flow solved numerically by Fromm. The practical stability criterion for the scheme is determined and a rigorous proof of convergence to a smooth solution is given. It is also shown how to construct an energy for any 3-level scalar difference equation.

1. Introduction

The investigation of stability of difference schemes for initial-value problems has largely amounted in practice to linearising the perturbation equations about a smooth solution, and setting the coefficients constant. The techniques of Fourier analysis are then applicable. This heuristic procedure is not always satisfactory since the effects of boundary conditions and nonlinearities may become important. The energy method can sometimes provide a rigorous analysis of these problems, and is complementary to Fourier analysis in the sense that the main difficulty in applying it is to construct a suitable energy for the constant coefficient scheme. Analysis of variable coefficient and nonlinear schemes may then be comparatively straightforward.

The present paper carries out this program, taking as an example the difference schemes used by Fromm [3] to calculate unsteady viscous flows past an obstacle in a channel. The construction of an energy for a given difference scheme is an unsolved problem in general, but a solution is given for a class of difference schemes which includes Fromm's. This scheme requires the vorticity to be specified on the obstacle and walls of the channel. It is shown that the boundary condition chosen by Fromm has no adverse effect on stability. Finally, the convergence of the finite difference solutions to sufficiently smooth solutions of the Navier-Stokes equations is established.

Most of the work presented in this paper is contained in a D. Phil. thesis submitted to the University of Oxford under the supervision of Dr. K. W. Morton, to whom I am indebted for many helpful discussions. I am grateful to the University of Edinburgh and the Science Research Council for providing financial support.

2. Statement of the Problem

The Navier-Stokes equations for two-dimensional viscous incompressible flow may be written in the form:

$$\begin{aligned} \frac{\partial \Omega}{\partial t} + \frac{\partial}{\partial x}(U\Omega) + \frac{\partial}{\partial y}(V\Omega) &= \nu \nabla^2 \Omega \\ \nabla^2 \Psi + \Omega &= 0 \end{aligned} \tag{2.1}$$

where ν is the kinematic viscosity, Ω the scalar vorticity, and Ψ is the stream function related to the cartesian components U and V of the fluid velocity by $U = \partial\Psi/\partial y$, $V = -\partial\Psi/\partial x$. The specific problem we shall consider is that of unsteady flow past a stationary obstacle in an infinite channel $0 \leqq y \leqq 1$. The boundary conditions are that the relative velocity between the fluid and the walls or obstacle is zero. It is assumed that the initial conditions can be adequately specified by some irrotational flow and an impulsive translatory motion of walls or fluid in the x-direction. We may then expect a wake to develop on the downstream side of the obstacle.

Fromm [3] has obtained a numerical solution of this problem by finite-difference methods. He assumes that the influence of the obstacle on the flow is negligible at points sufficiently far downstream, say at $x = L$. The flow variables at points on this line are assumed equal in value to those at corresponding points along a line upstream of the obstacle, say at $x = 0$. Thus we impose a periodic boundary condition in the x-direction. The main features of the problem are shown in Fig. 1. We denote by R the closed region outside and on the obstacle in $[0, L] \times [0, 1]$.

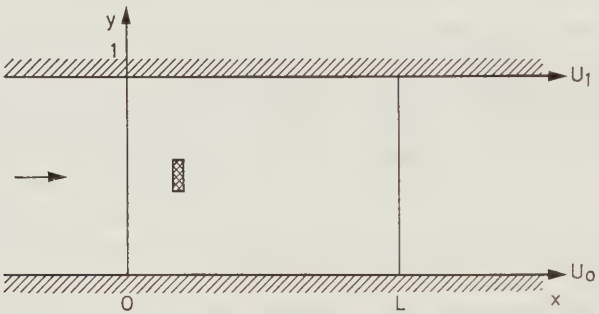

Fig. 1

It is assumed that the problem has a unique solution such that $\Omega, \Psi \in C^4(R)$ for $0 < t \leqq T$. If this assumption is not true, nonlinear instabilities may develop during the computation at points where the solution is varying rapidly, for instance round the leading edge of the obstacle. Such effects were indeed observed by Fromm in cases where the Reynolds number was high.

3. The Finite-Difference Equations

A square mesh with spacing h is set up over R and it is assumed that the boundaries of the obstacle pass through the mesh points. Let the mesh point i, j be at (ih, jh) for $0 \leqq i \leqq I = L/h$, $0 \leqq j \leqq J = 1/h$. We define the following operators:

(a) Translations: $T_{\pm x}\omega_{ij} \equiv \omega_{i\pm 1, j}$.

(b) Differences: $\Delta_{\pm x} \equiv \pm (T_{\pm x} - I)$,
$\Delta_{0x} \equiv \frac{1}{2}(\Delta_{+x} + \Delta_{-x})$.

(c) Averages: $\mu_{\pm x} \equiv \frac{1}{2}(T_{\pm x} + I)$,
$\mu_{0x} \equiv \frac{1}{2}(T_{+x} + T_{-x})$.

Similar operations are defined in the y-direction. In the difference equations we omit subscripts, thus $\omega_{ij}^n \sim \Omega(ih, jh, n\Delta t)$ is written as ω^n. ω and ψ are defined at mesh points and u, v as difference quotients of ψ; for convenience we shall choose to define them at mesh points also by

$$u^n = h^{-1}\Delta_{0y}\psi^n, \qquad v^n = -h^{-1}\Delta_{0x}\psi^n.$$

All mesh functions are regarded as elements of the real Hilbert space $\ell_2(R)$ endowed with inner product

$$(\phi, \chi) = \sum_R h^2 \phi_{ij}\chi_{ij} \quad \text{and} \quad \text{norm } \|\phi\| = \sqrt{(\phi, \phi)}.$$

We also define the discrete Fourier transform

$$\mathcal{F}(\phi) = \sum_R h^2 \phi_{ij} \exp[-ih(ik_1 + jk_2)].$$

Occasionally we shall deal with the space $\ell_\infty(R)$ with norm $|\phi| = \sup_R |\phi_{ij}|$. $\|\phi\|$ and $|\phi|$ are related by the inequalities $\|\phi\| \leqq |\phi|$, $|\phi| \leqq Kh^{-1}\|\phi\|$ where K is a constant depending on R.

The difference equations used by Fromm to solve (2.1) are the following:

$$\omega^{n+1} - \omega^{n-1} + 2\lambda(\Delta_{0x}(u\omega)^n + \Delta_{0y}(v\omega)^n) = 2\varrho(\overline{\mu_x + \mu_y}\,\omega^n - \omega^{n+1} - \omega^{n-1}) \quad (3.1\,\text{a})$$

and

$$\nabla^2\psi^n + \omega^n = 0, \quad (3.1\,\text{b})$$

where $\lambda = \Delta t/h$, $\varrho = 2\nu\,\Delta t/h^2$, and ∇^2 is the standard 5-point approximation to the Laplacian. In (3.1 a) an approximation of the leapfrog type is used for the transport terms, and the viscosity terms are approximated by a generalisation of the Dufort-Frankel scheme. Eqs. (3.1) are to be solved at all points in the interior of R. The boundary conditions are ψ constant along each wall and on the obstacle. We also require the following to hold:

for all i, $u_{i0} = U_0$ (velocity of the wall $y = 0$); $\quad v_{i0} = 0$;

and $\quad u_{iJ} = U_1$ (velocity of the wall $y = 1$); $\quad v_{iJ} = 0$;

also $u_{ij} = v_{ij} = 0$ for all points on the obstacle. In addition there are the periodicity conditions on ω and ψ: for all j, $\omega_{Ij} = \omega_{0j}$; $\psi_{Ij} = \psi_{0j}$. Finally, in order to solve (3.1 a) we must specify values of ω on the boundaries of R; these must be estimated by extrapolation, and may produce numerical instabilities. Fromm chose on physical grounds to set the boundary values equal to those at a distance h inwards along the normal; we shall show in § 5 that this choice does not affect stability. In § 4 we construct an energy for (3.1 a) with u and v constant, in § 5 we show that the practical stability criterion for (3.1 a) is $\lambda(|u| + |v|) < \frac{1}{2}$, and in § 6 we prove that solutions of (3.1) converge as $O(h^2)$ in ℓ_2 to a C^4 solution of (2.1) in R.

4. Analysis of the Constant-Coefficient Scheme

We need only consider the one-dimensional scheme

$$\omega^{n+1}-\omega^{n-1}+2\sigma\Delta_0\omega^n=\varrho(2\mu_0\omega^n-\omega^{n+1}-\omega^{n-1}) \tag{4.1}$$

since the energy for the two-dimensional problem will be an obvious generalisation. In [2] an energy for (4.1) was constructed in an intuitive manner from first principles. However this is not easy, and we give here a systematic solution of the problem which can be applied to any 3-level scalar difference equation.

We define the vector $\phi^n=[\omega^n,\omega^{n-1}]'$ and take the discrete Fourier transform $\mathscr{F}$ of (4.1). Using the relation

$$\mathscr{F}(T_\pm\omega)=e^{\pm i\xi}F(\omega)$$

where $\xi=kh$ we obtain $\mathscr{F}(\phi^{n+1})=G\mathscr{F}(\phi^n)$ where G is the amplification matrix.

Here

$$G=\begin{bmatrix}\alpha & \beta\\ 1 & 0\end{bmatrix}$$

where $\alpha=2(\varrho\cos\xi-i\sigma\sin\xi)/(1+\varrho)$, $\beta=(1-\varrho)/(1+\varrho)$.

The stability of (4.1) is defined in terms of G; we require $\|G^n\|$ to be uniformly bounded in ξ for all n. A necessary and sufficient condition for stability which we shall use was derived by Kreiss [4]. This is that there must exist a constant $C_H>0$ and for each ξ a positive definite Hermitean matrix $H(\xi)$ such that

$$C_H^{-1}I\leqq H\leqq C_H I \quad\text{and}\quad G^*HG\leqq H.$$

(By $A\leqq B$ we mean that $A-B$ is non-positive definite.)

Now the energy for (4.1) will be a quadratic form in ω^n, ω^{n-1} and powers of their translates $T_\pm\omega^n$, etc. This corresponds to a matrix H in Fourier space whose elements are polynomials in $e^{i\xi}$. It is an unsolved problem as to whether an H of this form exists for every stable scheme. However for matrices

$$G=\begin{bmatrix}\alpha & \beta\\ 1 & 0\end{bmatrix}$$

a suitable solution is proportional to

$$H=\begin{bmatrix}1 & -\frac{1}{2}\alpha\\ -\frac{1}{2}\bar{\alpha} & \frac{1}{2}(1+|\beta|^2)\end{bmatrix}.$$

Proof. We have to show that stability $\Rightarrow H\geqq C_H^{-1}I>0$ and $G^*HG\leqq H$. In terms of α and β these conditions become

$$2(1+|\beta|^2)-|\alpha|^2\geqq C>0 \tag{4.2}$$

and

$$|\beta|^2\leqq 1,\quad |\alpha+\bar{\alpha}\beta|\leqq 1-|\beta|^2. \tag{4.3}$$

First we determine the stability conditions on G. Application of the Kreiss-Buchanan criteria ([4, 1]) on the triangular form of G produces the following conditions on the eigenvalues λ_1 and λ_2:

$$|\lambda_1|, |\lambda_2| \leqq 1 \quad \text{(von Neumann condition)}$$

and

$$|\lambda_1| \cdot |1+\bar{\lambda}_1\lambda_2| \leqq M \max(|\lambda_1-\lambda_2|, 1-|\lambda_1|, 1-|\lambda_2|)$$

for some $M>0$.

Now, since the L.H.S. cannot vanish when there is a double root on the unit circle, the latter condition is equivalent to

$$\max(|\lambda_1-\lambda_2|, 1-|\lambda_1|, 1-|\lambda_2|) \geqq K>0. \tag{4.4}$$

Using the relations $\lambda_1\lambda_2=\beta$ and $2|\lambda_1|^2+2|\lambda_2|^2=|\alpha|^2+|\lambda_1-\lambda_2|^2$ we express (4.2) as

$$|\lambda_1-\lambda_2|^2+2(1-|\lambda_1|^2)(1-|\lambda_2|^2) \geqq C>0. \tag{4.5}$$

To show that (4.4) $\Rightarrow$ (4.5) we note that if the maximum of the L.H.S. of (4.4) is $|\lambda_1-\lambda_2|=K$ then (4.5) is satisfied with $C=K^2$. If on the other hand the maximum value is $1-|\lambda_1|=K$ then since $|\lambda_1-\lambda_2|^2 \geqq (|\lambda_1|-|\lambda_2|)^2$ the L.H.S. of (4.5) is greater than $f(K,\mu)=(K-\mu)^2+2K(2-K)\mu(2-\mu)$ where $\mu=1-|\lambda_2|$. From the fact that $f(K,\mu)$ has no stationary value with $0<\mu<1$ for fixed $K \leqq 1$, we can show that $f(K,1) \geqq f(K,\mu) \geqq f(K,0)=K^2$. Similarly for the case $1-|\lambda_2|=K$. Thus in all cases (4.4) implies (4.5) with $C=K^2$.

Finally, since $\lambda_1\lambda_2=\beta$ we must have $|\beta| \leqq 1$, and application of the von Neumann condition leads to

$$\tfrac{1}{2}|\alpha|^2+2|\beta+\tfrac{1}{4}\alpha^2| \leqq 1+|\beta|^2. \tag{4.6}$$

Isolating $|\beta+\frac{1}{4}\alpha^2|$ then squaring and expanding out, we obtain $|\alpha+\bar{\alpha}\beta| \leqq 1-|\beta|^2$ which is (4.3).

Remark. H is constructed as follows:

$$\text{Set } H=\begin{bmatrix} 1 & c \\ \bar{c} & b \end{bmatrix}$$

and then $G^*HG \leqq H$ if

$$|\beta|^2 \leqq b \quad \text{and} \quad |c-\beta\bar{c}-\bar{\alpha}\beta|^2 \leqq (b-|\beta|^2)(1-|\alpha|^2-2\mathscr{R}e\,\bar{\alpha}c-b).$$

b and c are chosen to make the R.H.S. a complete square; thus

$$c=-\tfrac{1}{2}\alpha, \qquad b=\tfrac{1}{2}(1+|\beta|^2).$$

Using this form for H, the corresponding energy for (4.1) is

$$\begin{aligned} S_n=&(1+\varrho^2)(\|\omega^n\|^2+\|\omega^{n-1}\|^2)+2\varrho\|\omega^n\|^2 \\ &-2\varrho(1+\varrho)(\omega^n, \mu_0\omega^{n-1})+2\sigma(1+\varrho)(\omega^n, \Delta_0\omega^{n-1}). \end{aligned}$$

The difference scheme is stable when $S_n \geqq C>0$ and $S_{n+1} \leqq S_n$. This in turn implies conditions on the mesh ratios ϱ and σ, which can be determined from

conditions (4.3). However they can be derived more directly from an analysis of the difference equations themselves. This involves some tedious algebra, but is essential in any case for the variable coefficient analysis.

First we define norms of diagonal differences

$$D_n = \tfrac{1}{2}\|\omega^n - T_+\,\omega^{n-1}\|^2 + \tfrac{1}{2}\|\omega^n - T_-\,\omega^{n-1}\|^2$$

and

$$F_n = \tfrac{1}{2}(1-\sigma)\,\|\omega^n - T_+\,\omega^{n-1}\|^2 + \tfrac{1}{2}(1+\sigma)\,\|\omega^n - T_-\,\omega^{n-1}\|^2.$$

Writing (4.1) as

$$\omega^{n+1} + \omega^{n-1} = 2\mu_0\omega^n - \varrho^{-1}(\omega^{n+1} - \omega^{n-1} + 2\sigma\Delta_0\omega^n) \tag{4.7}$$

and taking the inner product with $\omega^{n+1} - \omega^{n-1} + 2\sigma\Delta_0\omega^n$ we obtain

$$\begin{aligned}\|\omega^{n+1}\|^2 - \|\omega^{n-1}\|^2 + 2\sigma(\omega^{n+1} + \omega^{n-1},\, \Delta_0\omega^n)\\ = 2(\omega^{n+1} - \omega^{n-1},\, \mu_0\omega^n) + 4\sigma(\mu_0\omega^n,\, \Delta_0\omega^n)\\ - \varrho^{-1}\|\omega^{n+1} - \omega^{n-1} + 2\sigma\Delta_0\omega^n\|^2.\end{aligned} \tag{4.8}$$

Applying (A.6), (A.1) and (A.2) from the appendix we find that

$$F_{n+1} - F_n = -\varrho^{-1}\|\omega^{n+1} - \omega^{n-1} + 2\sigma\Delta_0\omega^n\|^2.$$

(Note that F_n is not a suitable energy since it is not equivalent to the ℓ_2-norm.)

By taking the inner product of (4.1) with $\omega^{n+1} + \omega^{n-1}$ we obtain a second identity

$$\begin{aligned}\|\omega^{n+1}\|^2 - \|\omega^{n-1}\|^2 = 2\varrho(\omega^{n+1} + \omega^{n-1},\, \mu_0\omega^n)\\ - 2\sigma(\omega^{n+1} + \omega^{n-1},\, \Delta_0\omega^n) - \varrho\|\omega^{n+1} + \omega^{n-1}\|^2.\end{aligned} \tag{4.9}$$

From (4.8) and (4.9) we obtain

$$\begin{aligned}S_{n+1} - S_n = {} & 2\varrho(\omega^{n+1} + \omega^{n-1},\, \mu_0\omega^n) - 2\sigma(\omega^{n+1} + \omega^{n-1},\, \Delta_0\omega^n)\\ & + 2\varrho(\omega^{n+1},\, \omega^{n+1} - \mu_0\omega^n) - 2\varrho(\omega^n,\, \omega^n - \mu_0\omega^{n-1})\\ & - \varrho\|\omega^{n+1} + \omega^{n-1}\|^2 - \varrho\|\omega^{n+1} - \omega^{n-1} + 2\sigma\Delta_0\omega^n\|^2\\ & + 2\sigma(1+\varrho)(\omega^{n+1} + \omega^{n-1},\, \Delta_0\omega^n)\\ & - 2\sigma\varrho^2(\omega^{n+1} + \omega^{n-1},\, \Delta_0\omega^n).\end{aligned}$$

By separating out the terms independent of σ and applying (A.1) we find after cancellation and rearrangement that

$$\begin{aligned}S_{n+1} - S_n = {} & -2\varrho[\|\omega^n\|^2 - 2(\omega^{n-1},\, \mu_0\omega^n) + \|\omega^{n-1}\|^2]\\ & - \varrho(\omega^{n+1} - \omega^{n-1} + 2\sigma\Delta_0\omega^n,\, 2\sigma\Delta_0\omega^n)\\ & - \varrho(\omega^{n+1} - \omega^{n-1},\, 2\sigma\Delta_0\omega^n)\\ & + 2\sigma\varrho(1-\varrho)(\omega^{n+1} + \omega^{n-1},\, \Delta_0\omega^n).\end{aligned}$$

Then, on substituting (4.1) in the product on the second line and using (A.1) we find that

$$\begin{aligned} S_{n+1} - S_n &= -2\varrho D_n + 4\sigma\varrho(\omega^{n-1}, \Delta_0\omega^n) \\ &= -2\varrho E_n \\ &\leqq 0 \quad \text{if} \quad |\sigma| \leqq 1. \end{aligned}$$

Finally, to show that S_n is positive definite we apply (A.1) and (A.6) and write it in the form

$$\begin{aligned} S_n = \|\omega^n + \sigma\Delta_0\omega^{n-1}\|^2 + \|\omega^{n+1}\|^2 - \sigma^2\|\Delta_0\omega^{n-1}\|^2 \\ + \varrho^2 D_n + \varrho(\omega^n + \sigma\Delta_0\omega^{n-1}, \omega^n - T_+\omega^{n-1} + \omega^n - T_-\omega^{n-1}). \end{aligned}$$

Estimation of the inner product by the Cauchy-Schwarz inequality and application of (A.4) leads to

$$\begin{aligned} S_n &\geqq \|\omega^n + \sigma\Delta_0\omega^{n-1}\|^2 + \varrho^2 D_n + (1-\sigma^2)\|\omega^{n-1}\|^2 \\ &\quad - \varrho[\varrho^{-1}\|\omega^n + \sigma\Delta_0\omega^{n-1}\|^2 + \varrho D_n] \\ &\geqq (1-\sigma^2)\|\omega^{n-1}\|^2. \end{aligned}$$

Therefore if $|\sigma| < 1$ we have

$$\|\omega^n\|^2 \leqq C S_{n+1} \leqq C S_n \leqq \cdots \leqq C S_0.$$

Since S_n is obviously bounded above in terms of the ℓ_2-norm, we have shown that the scheme is stable in ℓ_2 if $|\sigma| < 1$.

5. Variable-Coefficient Analysis and Boundary Conditions

In this section we derive the practical stability criterion for (3.1a) with u and v given and investigate the effects of boundary terms. We have to make the a priori assumption that u and v are Lipschitz continuous; this will be justified in § 6. First we obtain identities analogous to (4.8) and (4.9). (3.1a) can be written alternatively as

$$\begin{aligned} \omega^{n+1} + \omega^{n-1} &= \overline{\mu_x + \mu_y}\,\omega^n \\ &\quad - \tfrac{1}{2}\varrho^{-1}[\omega^{n+1} - \omega^{n-1} + 2\lambda(\Delta_{0x}(u\omega)^n + \Delta_{0y}(v\omega)^n)]. \end{aligned} \tag{5.1}$$

Then, taking the inner product of (5.1) with

$$\omega^{n+1} - \omega^{n-1} + 2\lambda(\Delta_{0x}(u\omega)^n + \Delta_{0y}(v\omega)^n)$$

leads to

$$\begin{aligned} &\|\omega^{n+1}\|^2 - \|\omega^{n-1}\|^2 + 2\lambda(\omega^{n+1} + \omega^{n-1}, \Delta_{0x}(u\omega)^n + \Delta_{0y}(v\omega)^n) \\ &\quad = (\omega^{n+1} - \omega^{n-1}, \overline{\mu_x + \mu_y}\,\omega^n) \\ &\qquad + 2\lambda(\Delta_{0x}(u\omega)^n + \Delta_{0y}(v\omega)^n, \overline{\mu_x + \mu_y}\,\omega^n) \\ &\qquad - \tfrac{1}{2}\varrho^{-1}\|\omega^{n+1} - \omega^{n-1} + 2\lambda(\Delta_{0x}(u\omega)^n + \Delta_{0y}(v\omega)^n)\|^2. \end{aligned} \tag{5.2}$$

On the other hand, taking the inner product of (3.1a) with $\omega^{n+1}+\omega^{n-1}$ yields

$$\begin{aligned}\|\omega^{n+1}\|^2-\|\omega^{n-1}\|^2+2\lambda\left(\omega^{n+1}+\omega^{n-1},\, \Delta_{0x}(u\omega)^n+\Delta_{0y}(v\omega)^n\right)\\ =2\varrho\,(\omega^{n+1}+\omega^{n-1},\, \overline{\mu_x+\mu_y}\,\omega^n)\\ -2\varrho\|\omega^{n+1}+\omega^{n-1}\|^2.\end{aligned}\tag{5.3}$$

Our object is to show that (3.1a) is stable in ℓ_2 if $\max(\lambda|u|,\, \lambda|v|)<\frac{1}{2}$. The proof is along the same lines as that for the one-dimensional constant-coefficient scheme, but we have to consider extra terms which arise from the variable coefficients, and values at the boundaries. The presence of the obstacle is neglected for the moment.

The first type of inner product to arise in (5.2) and (5.3) is

$$\begin{aligned}\left(\omega^{n-1},\, \Delta_{0x}(u\omega)^n\right)&=-\left(\omega^n,\, u^n\Delta_{0x}\omega^{n-1}\right)\\ &=-\left(\omega^n,\, u^{n-1}\Delta_{0x}\omega^{n-1}\right)+\left(\omega^n,\, \overline{u^{n-1}-u^n}\,\Delta_{0x}\omega^{n-1}\right)\\ &=-\left(\omega^n,\, \Delta_{0x}(u\omega)^{n-1}\right)+O(h)\,\|\omega^{n-1}\|\,\|\omega^n\|,\end{aligned}$$

by (A.7) and the periodicity condition on the vorticity.

Using (A.7) we also have

$$\begin{aligned}\left(\omega^{n-1},\, \Delta_{0y}(v\omega)^n\right)=&-\left(\omega^n,\, \Delta_{0y}(v\omega)^{n-1}\right)+O(h)\,\|\omega^{n-1}\|\,\|\omega^n\|\\ &+\tfrac{1}{2}\sum_{i=1}^{I-1}h^2\{\omega^{n-1}_{i,J-1}(v\omega)^n_{i,J}+\omega^{n-1}_{i,J}(v\omega)^n_{i,J-1}\\ &-\omega^{n-1}_{i,0}(v\omega)^n_{i,1}-\omega^{n-1}_{i,1}(v\omega)^n_{i,0}\}.\end{aligned}$$

Since $v^n_{i,0}=v^n_{i,J}=0$ for all i, n the sum reduces to

$$\tfrac{1}{2}\sum_{i=1}^{I-1}h^2\{\omega^{n-1}_{i,J}(v\omega)^n_{i,J-1}-\omega^{n-1}_{i,0}(v\omega)^n_{i,1}\}.$$

Now we shall show in § 6 that $|v-V|=O(h)$ in the interior of R. Since $V=0$ on the boundaries and its derivatives are bounded in R, $V=O(h)$ at mesh points adjacent to the boundaries and hence $v=O(h)$ also at these points. Thus the sum is at most $O(h)\,\|\omega^{n-1}\|\,\|\omega^n\|$.

The second type of inner product producing boundary terms is

$$\begin{aligned}(\mu_{0y}\omega^{n+1},\, \omega^n)=&(\omega^{n+1},\, \mu_{0y}\omega^n)\\ &+h^2/2\cdot\sum_{i=1}^{I-1}\{\omega^{n+1}_{i,J}\omega^n_{i,J-1}-\omega^{n+1}_{i,J-1}\omega^n_{i,J}-\omega^{n+1}_{i,1}\omega^n_{i,0}+\omega^{n+1}_{i,0}\omega^n_{i,1}\}\end{aligned}$$

on application of (A.1). If we define ω at the boundaries $j=0$ and $j=J$ by $\omega^n_{i,0}=\omega^n_{i,1}$ and $\omega^n_{i,J}=\omega^n_{i,J-1}$ for all n, we see that the above sum vanishes, so that this boundary condition has no adverse effect on stability. This is in fact the condition which was used by Fromm. It is evident that the above argument remains valid when the obstacle is included.

The third type of inner product we have to consider is $\left(\Delta_{0x}(u\omega)^n+\Delta_{0y}(v\omega)^n,\, \overline{\mu_x+\mu_y}\,\omega^n\right)$.

By (A.8) this is $O(h)\,\|\omega^n\|^2$ plus boundary terms in the y-direction. We have shown already that these can be neglected.

Let us now define the norm of diagonal differences

$$\begin{aligned} D_n &= \tfrac{1}{4}\|\omega^n - T_{+x}\omega^{n-1}\|^2 + \tfrac{1}{4}\|\omega^n - T_{-x}\omega^{n-1}\|^2 \\ &\quad + \tfrac{1}{4}\|\omega^n - T_{+y}\omega^{n-1}\|^2 + \tfrac{1}{4}\|\omega^n - T_{-y}\omega^{n-1}\|^2 \\ &= \|\omega^n\|^2 + \|\omega^{n-1}\|^2 - (\omega^n, \overline{\mu_x + \mu_y}\,\omega^{n-1}) \end{aligned}$$

if ω is defined as above at the boundaries, and the energy

$$\begin{aligned} S_n &= \|\omega^n\|^2 + \|\omega^{n-1}\|^2 + \varrho\,(\omega^n, 2\omega^n - \overline{\mu_x + \mu_y}\,\omega^{n-1}) \\ &\quad + \varrho^2 D_n + 2\lambda(1+\varrho)\left(\omega^n, \Delta_{0x}(u\omega)^{n-1} + \Delta_{0y}(v\omega)^{n-1}\right). \end{aligned}$$

When we calculate $S_{n+1} - S_n$ we perform the same manipulations as in the constant coefficient analysis and obtain the corresponding result, apart from additional terms which we have shown to be

$$\lambda O(h)\,\|\omega^n\|\,\|\omega^{n-1}\| = O(\Delta t)\,(\|\omega^n\|^2 + \|\omega^{n-1}\|^2)$$

when the correct boundary condition is used. We arrive at

$$\begin{aligned} S_{n+1} - S_n &= -2\varrho\left[D_n + 2\lambda\left(\omega^n, \Delta_{0x}(u\omega)^{n-1} + \Delta_{0y}(v\omega)^{n-1}\right)\right] \\ &\quad + O(\Delta t)\,(\|\omega^n\|^2 + \|\omega^{n-1}\|^2). \end{aligned}$$

Using the expanded form of D_n we split up the R.H.S. into terms involving translation operators in each dimension separately: Thus

$$\begin{aligned} S_{n+1} - S_n = -\varrho\,[&\tfrac{1}{2}(\omega^n - T_{+x}\omega^{n-1}, \overline{1 - 2\lambda T_{+x}}\,u^n \cdot \omega^n - T_{+x}\omega^{n-1}) \\ &+ \tfrac{1}{2}(\omega^n - T_{-x}\omega^{n-1}, \overline{1 + 2\lambda T_{-x}}\,u^n \cdot \omega^n - T_{-x}\omega^{n-1})] \\ &+ \lambda O(h)\,\|\omega^n\|\,\|\omega^{n-1}\| \\ &+ \text{corresponding terms in the } y\text{-direction}. \end{aligned}$$

The inner products are positive definite if $2\lambda \max(|u|, |v|) < 1$, and then $S_{n+1} - S_n \leqq O(\Delta t)\,(\|\omega^n\|^2 + \|\omega^{n-1}\|^2)$.

Finally we show that S_n is equivalent to the ℓ_2-norm, by the same argument as for the constant coefficient scheme. Using (A.8), we write

$$\begin{aligned} S_n &= \|\omega^n + \lambda\Delta_{0x}(u\omega)^{n-1} + \lambda\Delta_{0y}(v\omega)^{n-1}\|^2 \\ &\quad + \|\omega^{n-1}\|^2 - \lambda^2\|\Delta_{0x}(u\omega)^{n-1} + \Delta_{0y}(v\omega)^{n-1}\|^2 \\ &\quad + O(\Delta t)\,\|\omega^{n-1}\|^2 + \varrho^2 D_n \\ &\quad + \varrho\left(\omega^n + \lambda\Delta_{0x}(u\omega)^{n-1} + \lambda\Delta_{0y}(v\omega)^{n-1},\ 2\omega^n - \overline{\mu_x + \mu_y}\,\omega^{n-1}\right) \\ &\geqq \left(1 - O(\Delta t)\right)\|\omega^{n-1}\|^2 - \lambda^2\|\Delta_{0x}(u\omega)^{n-1} + \Delta_{0y}(v\omega)^{n-1}\|^2 \\ &\geqq \left(1 - O(\Delta t) - \lambda^2(|u| + |v|)^2\right)\|\omega^{n-1}\|^2 \end{aligned}$$

on estimation of the inner product in the same way as before. Hence if $\lambda \max(|u|, |v|) < \frac{1}{2}$ then a norm equivalent to the ℓ_2-norm has a growth which depends only on the Lipschitz constants of u and v. Thus the practical stability criterion for (3.1a) is $\lambda \max(|u|, |v|) < \frac{1}{2}$.

6. Convergence to a Smooth Solution

In this section we prove that the solutions of (3.1) converge as $O(h^2)$ in $\ell_2(R)$ to a C^4 solution of (2.1). First we obtain the error equations by substituting

$$\omega = \Omega + \omega'$$
$$u = U + u'$$
$$v = V + v'$$
$$\psi = \Psi + \psi'$$

in (3.1). We have

$$\omega'^{n+1} - \omega'^{n-1} + 2\lambda[\Delta_{0x}(u'\Omega + U\omega' + u'\omega')^n + \Delta_{0y}(v'\Omega + V\omega' + v'\omega')^n] = 2\varrho(\overline{\mu_x + \mu_y}\,\omega'^n - \omega'^{n+1} - \omega'^{n-1}) + \Delta t\, T_\omega, \tag{5.4a}$$

$$\nabla^2\psi'^n + \omega'^n = T_\psi \tag{5.4b}$$

where T_ω and T_ψ are the truncation errors of (3.1). It can be verified by Taylor expansions that if Ω, Ψ are C^4 in R,

$$|T_\omega| = O\big((\Delta t)^2 + h^2 + (\Delta t/h)^2\big)$$
$$|T_\psi| = O(h^2).$$

For a sequence of computations with $\Delta t/h^2$ fixed we have

$$|T_\omega| = O(h^2).$$

It is apparent that Eqs. (5.4) are of the same form as (3.1), with nonlinear and truncation error terms added on. Hence in the analysis of (5.4) we only need to consider the effect of these extra terms, since u and v in (3.1) are replaced in (5.4) by U and V, which are a fortiori Lipschitz continuous.

We intend to prove by induction that the ℓ_2-norm of the error $[\|\omega'^n\|^2 + \|u'^n\|^2 + \|v'^n\|^2]^{\frac{1}{2}}$ is $O(h^2)$. This is done by showing that (5.4) is stable in ℓ_2 under the induction hypothesis. We must estimate inner products of the form

$$\big(\omega'^{n+1} \pm \omega'^{n-1},\ \lambda\Delta_{0x}(u'^n\Omega^n + u'^n\omega'^n) + \lambda\Delta_{0y}(v'^n\Omega^n + v'^n\omega'^n) + \Delta t\, T_\omega\big).$$

The errors u' and v' are estimated as follows:

Defining

$$D_{0x} = h^{-1}\Delta_{0x}, \qquad D_{0y} = h^{-1}\Delta_{0y},$$

from

$$U + u' = D_{0y}(\Psi + \psi'); \qquad V + v' = -D_{0x}(\Psi + \psi')$$

we obtain

$$u' = D_{0y}\psi' + O(h^2): \quad v' = -D_{0x}\psi' + O(h^2).$$

Let α be the minimum eigenvalue of the difference operator $-\nabla^2$ in R. Its magnitude will depend on the size and shape of the obstacle, but will be quite large for a rectangular obstacle whose size and position are commensurable with the width of the channel. The Rayleigh quotient bound

$$-\frac{(\psi', \nabla^2\psi')}{(\psi', \psi')} \geqq \alpha$$

leads to the bound for ψ':

$$\|\psi'\|^2 \leqq \alpha^{-1}(\psi', -\nabla^2\psi') = -\alpha^{-1}(\psi', T_\psi - \omega') \leqq \alpha^{-1}\|\psi'\|\,\|T_\psi - \omega'\|.$$

Now take the inner product of (5.4b) with ψ'^n, apply (A.5) and we have

$$\begin{aligned}\|D_{+x}\psi'^n\|^2 + \|D_{+y}\psi'^n\|^2 &= -(\psi'^n, T_\psi - \omega'^n)\\ &\leqq \|\psi'^n\|\,\|T_\psi - \omega'^n\| \qquad (5.5)\\ &\leqq \alpha^{-1}(\|T_\psi\| + \|\omega'^n\|)^2.\end{aligned}$$

Thus $\|u'^n\|$ and $\|v'^n\|$ are $O(h^2)$ by hypothesis.

By the inequality between the ℓ_∞ and the ℓ_2-norms

$$|u'^n|, |v'^n| = O(h). \qquad (5.6)$$

Hence u and v are Lipschitz continuous and the assumption we made in § 5 is justified.

The additional inner products are expanded using the rule $\Delta_{0x}(u'\Omega) = \Delta_{0x}\Omega \cdot \mu_x u' + \mu_x\Omega \cdot \Delta_{0x}u'$, etc. to obtain

$$\begin{aligned}(\omega'^{n+1} \pm \omega'^{n-1},\ &\Delta_{0x}\Omega^n\mu_x u'^n + \Delta_{0y}\Omega^n\mu_y v'^n\\ &+ (\mu_x\Omega^n - \Omega^n)\Delta_{0x}u'^n + (\mu_y\Omega^n - \Omega^n)\Delta_{0y}v'^n\\ &+ \Omega^n(\Delta_{0x}u'^n + \Delta_{0y}v'^n)\\ &+ \Delta_{0x}(u'^n\omega'^n) + \Delta_{0y}(v'^n\omega'^n)\\ &+ \Delta t\, T_\omega).\end{aligned}$$

Since

$$\Delta_{0x}u^n + \Delta_{0y}v^n = 0 = \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y}$$

we can show by Taylor expansion that

$$\Delta_{0x}u'^n + \Delta_{0y}v'^n = \Delta_{0x}U^n + \Delta_{0y}V^n = O(h^3).$$

Hence on applying (5.6) we bound the inner product by

$$\begin{aligned}\lambda O(h)\,[\|\omega'^{n+1}\|^2 + \|\omega'^n\|^2 + \|\omega'^{n-1}\|^2 + \|u'^n\|^2 + \|v'^n\|^2]\\ + [\lambda O(h) + O(\Delta t)]\,(\|\omega'^{n+1}\|^2 + \|\omega'^{n-1}\|^2 + O(h^4)).\end{aligned}$$

Define

$$\begin{aligned}E_n = &\|\omega'^n\|^2 + \|\omega'^{n-1}\|^2 + \|u'^n\|^2 + \|v'^n\|^2\\ &+ \varrho^2 D_n' + 2\lambda(1+\varrho)\,(\omega'^n, \Delta_{0x}(U\omega')^n + \Delta_{0y}(V\omega')^n)\end{aligned}$$

where D_n' has the same form as D_n, in ω'. If $\lambda \max(|U|, |V|) < \frac{1}{2}$, and $E_n = O(h^4)$ we can show that E_n is equivalent to the ℓ_2 norm and

$$\begin{aligned}E_{n+1} - E_n &= O(\Delta t)\,(\|\omega'^n\|^2 + \|\omega'^{n-1}\|^2 + \|u'^n\|^2 + \|v'^n\|^2)\\ &\quad + O(\Delta t)\,(\|\omega'^{n+1}\|^2 + \|\omega'^{n-1}\|^2 + O(h^4))\\ &= O(\Delta t)\,(E_{n+1} + E_n) + \Delta t\, O(h^4).\end{aligned}$$

Hence constants L and M exist such that for sufficiently small Δt,

$$\frac{E_{n+1}-E_n}{\Delta t} \leqq L E_n + M h^4$$

and thus

$$\begin{aligned} E_n &\leqq e^{LT}[E_0 + L^{-1} M h^4] \\ &= O(h^4) \quad \text{for all } n \text{ such that } \quad n\Delta t \leqq T, \end{aligned}$$

given that the initial error E_0 is $O(h^4)$; i.e. $O(h^2)$ in ℓ_2. Our induction assumption is thus justified, and the errors about a C^4 solution of (2.1) are $O(h^2)$ in ℓ_2, and hence $O(h)$ pointwise in the interior of R.

Appendix

Let u and v be functions defined on a mesh with spacing h in $[0, L] \times [0, 1]$ such that $u_{ij} = u(ih, jh)$ for $0 \leqq i \leqq I = L/h$, $0 \leqq j \leqq J = 1/h$. The following relations can then be proved:

In the x-direction we have

$$(u, \mu_{0x} v) = (\mu_{0x} u, v) + h/2\,(u_{I-1,j} v_{Ij} - u_{Ij} v_{I-1,j} - u_{0j} v_{1j} + u_{1j} v_{0j}) \tag{A.1}$$

with the corresponding result in the y-direction.

Next,

$$(u, \Delta_{0x} v) = -(\Delta_{0x} u, v) + h/2\,(u_{I-1,j} v_{Ij} + u_{Ij} v_{I-1,j} - u_{0j} v_{1j} + u_{1j} v_{0j}). \tag{A.2}$$

When u and v are periodic in x, $(u, \Delta_{0x} v) = -(\Delta_{0x} u, v)$ and in particular

$$(u, \Delta_{0x} u) = 0. \tag{A.3}$$

Also in the periodic case $\|T_{\pm x} u\|^2 = \|u\|^2$ and hence

$$\|\Delta_{0x} u\|^2 \leqq \tfrac{1}{4}(\|T_{+x} u\|^2 + \|T_{-x} u\|^2) = \|u\|^2. \tag{A.4}$$

If u is constant at the boundaries

$$(u, \delta_x^2 u) = -\|\Delta_{+x} u\|^2. \tag{A.5}$$

In two dimensions we have

$$(\overline{\mu_x + \mu_y} u, \overline{\Delta_{0x} + \Delta_{0y}} u) = 0. \tag{A.6}$$

These identities are modified when the inner product involves a variable coefficient a, which we assume to be Lipschitz continuous. Using the identities

$$\begin{aligned} \Delta_{0x}(a u) &= a \Delta_{0x} u + \tfrac{1}{2}(T_{+x} u \Delta_{+x} a + T_{-x} u \Delta_{-x} a) \\ \mu_{0x}(a u) &= a \mu_{0x} u + \tfrac{1}{2}(T_{+x} u \Delta_{+x} a - T_{-x} u \Delta_{-x} a) \end{aligned}$$

we find that (A.2) generalises to

$$\begin{aligned} (u, \Delta_{0x}(a v)) = &-(v, \Delta_{0x}(a u)) + O(h)\,\|u\|\,\|v\| \\ &+ h/2\big(u_{I-1,j}(a v)_{Ij} + u_{Ij}(a v)_{I-1,j} - u_{0j}(a v)_{1j} - u_{1j}(a v)_{0j}\big) \end{aligned} \tag{A.7}$$

and (A.6) becomes

$$(\overline{\mu_x + \mu_y} u, \overline{\Delta_{0x} + \Delta_{0y}}(a u)) = O(h)\,\|u\|^2. \tag{A.8}$$

References

1. Buchanan, M. L.: A necessary and sufficient condition for stability of difference schemes for initial-value problems. J. Soc. Indust. and Appl. Maths. **11**, 919—935 (1963).
2. Campbell, N. G.: Stability theory of finite-difference schemes for partial differential equations. Doctoral thesis, Univ. of Oxford, Oxford, 1968.
3. Fromm, J. E.: The time-dependent flow of an incompressible viscous fluid, in: Methods of computational physics, v. 3, pp. 346—382. New York: Academic Press 1964.
4. Kreiss, H. O.: Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren. Nordisk Tidskrift Informations-Behandling **2**, 153—181 (1962).

N. G. Campbell
Atlas Computer Laboratory
Chilton, Didcot, Berkshire
England

Numer. Math. 14, 448–467 (1970)

Nichtnegativitäts- und substanzerhaltende Differenzenschemata für lineare Diffusionsgleichungen*

Rudolf Gorenflo

Eingegangen am 25. Juli 1969

Abstract. Explicit difference schemes preserving non-negativity and mass as does the described diffusion process are given for the Fokker-Planck partial differential equation (Kolmogorov's forward differential equation) in $\mathbb{R}^n$. These schemes can be interpreted as descriptions of discrete non-stationary inhomogeneous random walks. They are stable in the maximum norm, and if an initial condition is given, they allow approximate solution of the differential equation as well as approximate Monte-Carlo simulation of the basic diffusion process. Essential conditions are the boundedness of the coefficients of the differential equation and of some of their derivatives and a strong diagonal dominance of the diffusion matrix. The results can be generalized to general linear parabolic differential equations (conservation of mass being lost, of course). By means of the developed method it is also possible to construct difference approximations of non-negative type to linear elliptic operators in $\mathbb{R}^n$ if these operators satisfy analogous conditions.

1. Einleitung

Diese Arbeit behandelt nichtnegativitäts- und substanzerhaltende explizite Differenzenschemata für die allgemeine Fokker-Planck-Gleichung im $\mathbb{R}^n$, $n \geqq 1$,

$$\text{(FP)} \qquad \frac{\partial u}{\partial t} = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial^2}{\partial x_j \, \partial x_k} (b_{jk} u) - \sum_{j=1}^{n} \frac{\partial}{\partial x_j} (a_j u)$$

und für die Fokker-Planck-Gleichung mit selbstadjungiertem elliptischem Anteil[1]

$$\text{(FPS)} \qquad \frac{\partial u}{\partial t} = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial}{\partial x_j} \left(b_{jk} \frac{\partial u}{\partial x_k} \right).$$

In beiden Fällen sei eine Anfangsbedingung

$$\text{(A)} \qquad u(x, 0) = g(x), \qquad x \in \mathbb{R}^n,$$

gegeben, und man interessiert sich für die Lösung $u(x, t)$ bzw. eine Näherungslösung in $\mathfrak{D} = \mathbb{R}^n \times \{t \mid 0 \leqq t \leqq T\}$ mit einem festen $T > 0$. Ein Punkt des $\mathbb{R}^n$ mit den Koordinaten $x_1, x_2, \ldots, x_n$ wird der Kürze halber mit x bezeichnet. Im

* Diese Arbeit entstand im Rahmen des Vertrags zwischen dem Institut für Plasmaphysik (GmbH) und der Europäischen Atomgemeinschaft über die Zusammenarbeit auf dem Gebiet der Plasmaphysik.

1 Mit $a_j = \frac{1}{2} \sum_{k=1}^{n} \frac{\partial b_{jk}}{\partial x_k}$ geht (FPS) in (FP) über.

folgenden sei stets $(x, t) \in \mathfrak{D}$. Alle auftretenden Variablen und Funktionen seien reell.

Die Matrix $B(x, t) = \{b_{jk}(x, t)\}$ *sei symmetrisch (also* $b_{jk} = b_{kj}$*) und gleichmäßig positiv definit, so daß* (FP) *und* (FPS) *gleichmäßig parabolische Differentialgleichungen sind.* B ist gleichmäßig positiv definit, wenn der kleinste Eigenwert von B nie kleiner wird als eine feste positive Zahl und wenn alle b_{jk} beschränkt sind.

Wir müssen noch das Erfülltsein gewisser Glattheitsbedingungen für die Koeffizienten $b_{jk}(x, t)$, $a_j(x, t)$ und die Funktion $g(x)$ fordern, damit die Lösung $u(x, t)$ eindeutig existiert und ebenfalls so glatt ist, wie es für die in den nächsten Kapiteln durchzuführenden Konsistenz- und Stabilitätsüberlegungen notwendig ist. Aussagen über die eindeutige Existenz der Lösung und den Zusammenhang der Glattheit der Koeffizienten und der Anfangsbedingung mit der Glattheit der Lösung findet man in [6].

Für unsere Abschätzungen benötigen wir die Existenz und Beschränktheit der b_{jk} *und gewisser ihrer Ableitungen nach* x *bis zur Ordnung* 4, *der* a_{jk} *und gewisser ihrer Ableitungen nach* x *bis zur Ordnung* 3, *der Lösung* u *und gewisser ihrer Ableitungen nach* x *bis zur Ordnung* 4 *und ihrer Ableitungen nach* t *bis zur Ordnung* 2 *(hierzu müssen natürlich auch die* b_{jk} *und die* a_j *bezüglich* t *genügend glatt sein).*

Wir verzichten darauf, möglichst schwache Voraussetzungen zu formulieren, damit diese Erfordernisse auch für die Lösung u erfüllt sind, sondern begnügen uns mit der Feststellung, daß sie erfüllbar sind. (Man vgl. Kapitel 3 des Buches [6].) *Wir nehmen sie im folgenden stets als erfüllt an, ohne dies immer ausdrücklich zu erwähnen. Wir verwenden gelegentlich den etwas vagen Ausdruck „hinreichend glatt", der die Existenz, Beschränktheit und Stetigkeit der jeweils benötigten Ableitungen postulieren soll.*

Speziell $g(x)$ *sei beschränkt und habe beschränkte Ableitungen nach* x *bis zur Ordnung* 4. *Wenn man dann fordert, daß* $u(x, t)$ *beschränkt ist, so sind die Cauchyschen Anfangswertprobleme* (FP, A) *und* (FPS, A) *in der Maximum-Norm korrekt gestellt.*

(FP) und (FPS) beschreiben die Diffusion einer im $\mathbb{R}^n$ verteilten Substanz, wenn man $u(x, t)$ als die Dichte dieser Substanz am Orte x zur Zeit t auffaßt. Es bietet sich deshalb der Versuch an, zwei wesentliche Eigenschaften dieses Diffusionsprozesses, nämlich die Nichtnegativitätserhaltung[2] und die Substanzerhaltung[2] durch ein explizites Differenzenschema zu imitieren. Dies führt auf Differenzenapproximationen nichtnegativen Typs, die auch unabhängig von der angedeuteten physikalischen Analogie von Interesse sind. Man erhält nämlich gleichzeitig Approximationen nichtnegativen Typs für die auf der rechten Seite von (FP) und (FPS) stehenden elliptischen Differentialausdrücke. Solche Approximationen sind nützlich bei der numerischen Behandlung nichtlinearer elliptischer Differentialgleichungen und bei Monte-Carlo-Methoden für lineare elliptische Randwertprobleme (man vgl. [1, 2, 4, 8, 12]).

2 Nichtnegativitätserhaltung bedeutet: Ist $u(x, s) \geqq 0$, so ist auch $u(x, t) \geqq 0$ für $t > s$. Man vgl. [6], S. 42ff. Substanzerhaltung bedeutet: Die Substanzänderung in einem Gebiet $V \subset \mathbb{R}^n$ mit glattem Rand kommt dadurch zustande, daß über diesen Rand Substanz zufließt oder abfließt. Die Substanzerhaltungseigenschaft läßt sich leicht mittels des Gaußschen Integralsatzes einsehen.

Faßt man $u(x, t)$ als Wahrscheinlichkeitsdichte des Aufenthalts eines irrenden Teilchens am Orte x zur Zeit t auf, so beschreiben (FP) und (FPS) stetige Markov-Prozesse[3] (man vgl. [3]). Die gegebenenfalls existierenden nichtnegativitäts- und substanzerhaltenden Differenzenapproximationen können dann interpretiert werden als Beschreibung diskreter Irrfahrten, und man kann sie Differenzenschemata vom Irrfahrttypus nennen ([7]). Sie gestatten also neben der approximativen numerischen Lösung des Anfangswertproblems auch die approximative Monte-Carlo-Simulation des Diffusionsprozesses. Im konkreten Einzelfall können natürlich bezüglich Ausnutzung eines Computers effektivere Methoden vorhanden sein[4]. Im Falle $n = 1$ sind einige einfache Beispiele diskreter Irrfahrten und ihr Übergang bei verschwindender Schrittweite in Diffusionsgleichungen schon lange bekannt (man vgl. z.B. [5], S. 354ff., [11, 13]). Für den Fall $n \geqq 2$ mit nur von t abhängenden b_{jk} und a_j vgl. man [9].

In diesem Zusammenhang sei auf die wahrscheinlichkeitstheoretische Bedeutung von B und a hingewiesen. Bezeichnet $x = x(t)$ die Bahn eines gemäß dem durch (FP) beschriebenen Markov-Prozeß irrenden Teilchens, und ist $\langle \cdot \rangle$ der Operator der mathematischen Erwartung, ist ferner $\Delta t > 0$ und $\Delta x = x(t + \Delta t) - x(t)$, so ist für $\Delta t \to 0$

$$\langle \Delta x \rangle = a(x, t)\, \Delta t + o(\Delta t),$$

$$\langle \Delta x_j \Delta x_k \rangle = b_{jk}(x, t)\, \Delta t + o(\Delta t).$$

Man nennt $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ den Driftvektor, $B = (b_{jk})$ die Diffusionsmatrix.

Theorem 2 von Motzkin und Wasow ([12]) läßt hoffen, unter ziemlich allgemeinen Bedingungen Differenzenschemata der gewünschten Art finden zu können; beschränkt man sich aber auf ein von vornherein gegebenes System von Nachbargitterpunkten, so lehrt Theorem 1 von [12], daß man ziemlich einschränkende Bedingungen an die elliptischen Differentialausdrücke auf den rechten Seiten stellen muß. In dieser Arbeit sollen Schemata einer noch zu präzisierenden einfachen Struktur — und hinreichende Bedingungen für ihre Existenz — angegeben werden. Wir werden uns beschränken auf kubische Gitternetze im $\mathbb{R}^n$ mit den Gitterpunkten $x = m h$, $h > 0$, m ein Multiindex, und auf das Nachbarschaftssystem von x, das in diesem Netz alle Punkte mit den Abständen 0, h und $\sqrt{2}h$

3 Man nennt dann (FP) auch die Vorwärtsdifferentialgleichung von Kolmogorov.

4 Bekanntlich ist ja auch die Nichtnegativität der Koeffizienten eines expliziten Schemas für parabolische Differentialgleichungen keine notwendige Bedingung für Stabilität ([4], S. 113ff.). Man wird auch, falls nur ein endliches Gebiet vorliegt und Randbedingungen gegeben sind, mit Vorteil implizite Methoden benutzen. — Eigentlicher Anlaß für die Entstehung dieser Arbeit war eine Vortragsreihe über Diffusionsprozesse, die der Verfasser vor Physikern des Instituts für Plasmaphysik hielt. Er beschäftigte sich in diesem Zusammenhang mit approximierenden diskreten Irrfahrtprozessen und stellte sich die Frage, wie weit man die in [5] für $n = 1$ beschriebenen Irrfahrt-Approximationen verallgemeinern kann. Da er sich gleichzeitig in die Theorie der Differenzenschemata einarbeitete, benutzte er den Apparat dieser Theorie zur Untersuchung von Konvergenz und Stabilität. Er hofft, den Nutzen der beschriebenen Schemata bei numerischen Rechnungen und zweckmäßige Modifikationen für parabolische Randwertprobleme in einer späteren Arbeit untersuchen und diskutieren zu können.

von x enthält. Als wesentliche Bedingung für die Existenz entsprechender Differenzenschemata wird sich die Diagonaldominanz der Diffusionsmatrix B ergeben. Mit dieser Bedingung ist eine wichtige Klasse von in den Anwendungen wichtigen linearen Diffusionsgleichungen erfaßt.

2. Konsistenz und Stabilität des Differenzenschemas, auch für allgemeinere lineare parabolische Differentialgleichungen

Wir geben der Vollständigkeit halber einen kurzen Abriß (in Anlehnung an [10], Kapitel 9) der Theorie von Konsistenz, Stabilität und Konvergenz für das inhomogene Anfangswertproblem

$$\frac{\partial u}{\partial t} = A\,u + f, \qquad u(x,0) = g(x), \qquad f = f(x,t), \tag{1}$$
$$u = u(x,t) \quad \text{gesucht in } 0 \leqq t \leqq T.$$

Dieser Abriß ist zugeschnitten auf unsere Bedürfnisse, enthält aber im wesentlichen wohlbekannte Aussagen. $A = A(x,t)$ sei ein linearer Differentialoperator, der nur Ableitungen nach den Komponenten x_j von x enthalte (auch Ableitungen höherer Ordnung sind zulässig). Das Problem (1) sei korrekt gestellt in der Maximum-Norm, d. h., u soll eindeutig existieren und normstetig von den Anfangswerten abhängen — in einem geeigneten Funktionenraum. Für irgendeine beschränkte Funktion $w(x,t)$ ist die Norm so definiert:

$$\|w(t)\| = \sup_{x \in \mathbb{R}^n} |w(x,t)|.$$

Wir diskretisieren so: Zu $h > 0$ und $\tau > 0$ mit $\tau \leqq \tau_0(h)$, $h \leqq h_0$, $\tau_0(h)$ eine geeignete Funktion von h, betrachten wir die Punkte $x = m\,h$, $t = \nu\,\tau$ des Gitters $D = D(h,\tau)$. Dabei ist m ein Multiindex mit den n ganzzahligen Komponenten m_j, $-\infty < m_j < \infty$, $j = 1, 2, \ldots, n$, und ν durchläuft die ganzen Zahlen aus $0 \leqq \nu \leqq T/\tau$. Es interessiert der Grenzübergang $h \to +0$.

Setzt man $L = \frac{\partial}{\partial t} - A$, so geht (1) über in

$$Lu = f, \qquad u(x,0) = g(x), \qquad u \text{ gesucht in } 0 \leqq t \leqq T. \tag{2}$$

Konsistenz einer Differenzenapproximation mit dem Gitter D

$$L_D U = F, \qquad U(x,0) = G(x), \qquad U \text{ gesucht in } D = D(h,\tau) \tag{2'}$$

zu (2) bedeutet nun, daß für die lokalen Abbruchfehler[5] $\lambda, \tilde{\varphi}, \tilde{\gamma}$ in

$$L_D u = f + \lambda, \qquad F = f + \tilde{\varphi}, \qquad G = g + \tilde{\gamma} \tag{3}$$

bei $h \to 0$ gilt

$$\sup_{0 \leqq t \leqq T} \|\lambda(t)\|_D \to 0, \qquad \sup_{0 \leqq t \leqq T} \|\tilde{\varphi}(t)\|_D \to 0, \qquad \sup_{0 \leqq t \leqq T} \|\tilde{\gamma}(t)\|_D \to 0. \tag{4}$$

5 Es ist zweckmäßig, hier Funktionen F und G einzuführen, da es im Sinne der diskreten Imitation der zugrunde liegenden Diffusionsprozesse liegt, anstelle von f und g Mittelwerte über gewisse mit dem Gitter zusammenhängende Würfelbereiche zu nehmen.

Man kann auch so sagen: Der Differenzenoperator L_D ist eine $O(h^p)$-konsistente Approximation an den Differentialoperator L, wenn für hinreichend glatte Funktionen $v(x,t)$ gilt

$$\|(L_D - L)\, v\|_D = O(h^p).$$

Dabei sei $p > 0$. Gewünscht ist, daß bei $h \to 0$ die Lösung U von (2') gegen die Lösung u von (2) strebt. Wir müssen noch die Gitternorm $\|\cdot\|_D$ definieren. Für irgendeine beschränkte Funktion $w(x,t)$ ist

$$\|w(t)\|_D = \|w(t)\|_{D(h,\tau)} = \sup_{x \in \mathbb{R}^n \cap D(h,\tau)} |w(x,t)|.$$

Nun wird in der wirklichen Rechenpraxis nicht U aus (2') ausgerechnet, sondern es wird, mit lokalen Rundungsfehlern φ und γ, eine Gitterfunktion W aus dem Problem

$$(2'') \qquad L_D W = F + \varphi, \quad W(x,0) = G(x) + \gamma, \quad W \text{ gesucht in } D,$$

ausgerechnet. Es gilt das

Stabilitätslemma 1. *Wenn für beliebige beschränkte Funktionen F und G aus (2') folgt, daß mit positiven Konstanten K_1, K_2, die von den Inhomogenitäten F, G und von der Feinheit des Gitters $D = D(h,\tau)$ unabhängig sind,*

$$(5) \qquad \|U(t)\|_D \leqq K_1 \sup_{0 \leqq s \leqq T} \|F(s)\|_D + K_2 \|G\|_D, \qquad t = \nu\tau,$$

gilt, so gilt für die Lösung W des Problems (2'') die Ungleichung

$$(6) \qquad \|W(t) - u(t)\|_D \leqq K_1 \sup_{0 \leqq s \leqq T} \|\tilde{\varphi} + \varphi - \lambda\|_D + K_2 \|\tilde{\gamma} + \gamma\|_D.$$

Das Lemma lehrt, daß bei rundungsfehlerfreier Rechnung $W(x,t)$ gleichmäßig gegen $u(x,t)$ konvergieren würde, und daß die auftretenden Rundungsfehler, deren Größenordnung man in der Praxis oft schätzen kann, nicht beliebig stark vergrößert werden, wenn (5) gilt. (5) ist die definierende Bedingung für Stabilität des Verfahrens.

Zum Beweis des Lemmas braucht man nur zu beachten, daß $L_D(W - u) = \tilde{\varphi} + \varphi - \lambda$ und $W(x,0) - u(x,0) = \tilde{\gamma} + \gamma$ ist.

Wir formulieren noch als hinreichende Bedingung für die Gültigkeit von (5) das bekannte „Indexkriterium" als

Stabilitätslemma 2. *Wenn mit einer von der Feinheit des Gitters unabhängigen Konstanten $C \geqq 0$ aus (2')*

$$(7) \qquad \|U(t+\tau)\|_D \leqq (1 + C\tau)\, \|U(t)\|_D + \tau \|F(t)\|_D, \qquad t = \nu\tau,$$

folgt, so gilt (5) mit

$$K_1 = (e^{CT} - 1)/C \quad \text{für} \quad C > 0, \qquad K_1 = T \quad \text{für} \quad C = 0, \qquad K_2 = e^{CT}.$$

Mit etwas handlicherer Bezeichnungsweise folgt dieses Lemma sofort aus folgender Aussage: Es seien $\{s_\nu\}$ und $\{v_\nu\}$ Folgen nichtnegativer Zahlen mit $v_\nu \leqq (1 + C\tau) v_{\nu-1} + \tau s_{\nu-1}$. Ferner sei $C \geqq 0$, $N = [T/\tau]$, $0 \leqq \nu \leqq N$, $S = \max_{0 \leqq \nu \leqq N-1} s_\nu$. Dann ist $v_\nu \leqq K_2 v_0 + K_1 S$. Diese letzte Ungleichung ergibt sich so: Eine einfache Re-

kursion und anschließende Abschätzung liefert

$$v_\nu \leqq (1+C\tau)^\nu v_0 + \tau S \sum_{j=1}^{\nu-1} (1+C\tau)^{j-1}$$

$$\leqq (1+C\tau)^{T/\tau} v_0 + \tau S \sum_{j=1}^{N} (1+C\tau)^{j-1} \leqq K_2 v_0 + K_1 S.$$

Im folgenden werden wir das hinreichende Stabilitätslemma 2 benutzen. Jetzt wenden wir uns unseren Diffusionsgleichungen zu. (FP) und (FPS) lassen sich durch Ausdifferenzieren umformen in eine lineare parabolische Differentialgleichung der Gestalt

$$\frac{\partial u}{\partial t} = P u \quad \text{mit} \quad P u = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} b_{jk} \frac{\partial^2 u}{\partial x_j \partial x_k} + \sum_{j=1}^{n} \tilde{a}_j \frac{\partial u}{\partial x_j} + \tilde{c}\, u. \tag{8}$$

Für die Diskretisierung D sei, mit später noch weiter einzuschränkendem konstantem $\sigma > 0$,

$$\tau = 2\sigma h^2. \tag{9}$$

Das Differenzenschema, nach $U(x, t+\tau)$ aufgelöst, lautet

$$U(x, t+\tau) = \sum_{r \in \mathfrak{R}} p_r(x, t; h, \tau)\, U(x + rh, t), \quad t = \nu\tau, \quad x = mh,$$
$$\nu = 0, 1, 2, \ldots, [T/\tau] - 1. \tag{10}$$

Dabei sei $\mathfrak{R}$ ein festes System von Multiindices r mit $r' = (r_1, r_2, \ldots, r_n)$. Setzt man, um den Anschluß an die weiter oben dargestellte Konsistenz- und Stabilitätstheorie zu finden, $A = P$ und $L = \frac{\partial}{\partial t} - P$, so ist für eine hinreichend glatte Funktion $v(x, t)$ sinngemäß

$$L_D v = \frac{1}{\tau} \{v(x, t+\tau) - v(x, t)\} - \frac{1}{\tau} \sum_{r \in \mathfrak{R}} p_r(x, t; h, \tau)\, v(x + rh, t) + \frac{v(x, t)}{\tau}.$$

Wegen

$$\frac{1}{\tau} \{v(x, t+\tau) - v(x, t)\} = \frac{\partial v}{\partial t} + O(\tau)$$

ergibt sich als $o(1)$-Konsistenzbedingung

$$(P_D - P) v = o(1).$$

Da diese für jede hinreichend glatte Funktion v erfüllt sein muß, ergibt sich durch Taylor-Entwicklung ihre Äquivalenz mit folgenden 3 Bedingungen (dabei ist $p_r = p_r(x, t; h, \tau)$):

$$\left.\begin{aligned} &\text{(K 1)} \quad \frac{1}{\tau}\left(\sum_{r \in \mathfrak{R}} p_r - 1\right) \to \tilde{c} \\ &\text{(K 2)} \quad \frac{h}{\tau} \sum_{r \in \mathfrak{R}} r_j p_r \to \tilde{a}_j \\ &\text{(K 3)} \quad \frac{h^2}{\tau} \sum_{r \in \mathfrak{R}} r_j r_k p_r \to b_{jk} \end{aligned}\right\} \text{ bei } h \to 0,$$

bzw.

$$\sum_{r\in\mathfrak{R}} p_r = 1+\tilde{c}\,\tau+o(\tau), \quad \sum_{r\in\mathfrak{R}} r_j p_r = 2\sigma h\tilde{a}_j+o(h),$$

$$\sum_{r\in\mathfrak{R}} r_j r_k p_r = 2\sigma b_{jk}+o(1).$$

Notwendig und hinreichend für Nichtnegativitäts-Erhaltung ist das Erfülltsein der Bedingung

(NN) $$p_r(x,t;h,\tau)\geqq 0, \quad r\in\mathfrak{R},$$

die besagt, daß die Approximation von nichtnegativem Typ ist.

Wir haben jetzt die Mittel bereitgestellt, einen allgemeinen fast trivialen Satz über lineare parabolische Differentialgleichungen und ihre Approximation durch Differenzenschemata nichtnegativen Typs auszusprechen.

Satz 1. *Das lineare gleichmäßig parabolische Anfangswertproblem*

(8′) $$\frac{\partial u}{\partial t} = Pu+f \quad \textit{mit } P \textit{ wie in } (8), \quad u(x,0)=g(x),$$

hinreichend glatten Koeffizienten, hinreichend glatter Funktion $f=f(x,t)$ und hinreichend glatter Anfangsbedingung $g(x)$ werde approximiert durch ein Schema

(10′) $$U(x,t+\tau)=\sum_{r\in\mathfrak{R}} p_r(x,t;h,\tau)\,U(x+rh,t)+\tau F(x,t)$$

mit $U(x,0)=G(x)$. Gesucht ist eine Näherungslösung in $0\leqq t\leqq T$. F und G seien konsistente Approximationen an f und g, (10′) *erfülle die Konsistenzbedingungen* (K1), (K2), (K3) *und die Nichtnegativitäts-Bedingung* (NN). *Dann ist* (10′) *ein stabiles Schema*[6]*, das zu* (8′) *konsistent ist.*

Beweis. Aus (10′), (8′) und (NN) folgt

$$\|U(t+\tau)\|_D \leqq \Big(\sum_r p_r\Big)\,\|U(t)\|_D+\tau\|F(t)\|_D.$$

Mit (K1) ergibt sich nun sofort (7). Zu $\varepsilon>0$ gibt es ein $h_0>0$ so, daß man $C=\sup_{0\leqq t\leqq T}|\tilde{c}|+\varepsilon$ für $h\leqq h_0$ nehmen kann.

In Satz 1 ist es nicht erforderlich, daß die verkürzte Gleichung $\partial u/\partial t = Pu$ eine Fokker-Planck-Gleichung ist (B muß natürlich symmetrisch und gleichmäßig positiv definit sein). Die allgemeine Gl. (8′) entsteht, wie leicht nachzurechnen ist, aus einer Gl. (FP) durch additive Hinzufügung eines Terms $c(x,t)u$ und eines Quellterms $f(x,t)$. Die Interpretation dieser Terme als Entstehungsdichte von Substanz liegt auf der Hand. $c\,u$ bedeutet Entstehungsdichte proportional zur gerade vorhandenen Substanzdichte. (Ein negativer Term $c\,u$ oder f entspricht einer Vernichtungsdichte von Substanz von der Größe seines Betrags.) Hat man zu (FP) eine Approximation, in der $p_0(x,t;h,\tau)$ nicht kleiner als eine positive Konstante ist (dies werden wir später durch hinreichend kleine Wahl von σ erzwingen können), so ist für hinreichend kleines h (und damit hinreichend kleines τ) auch $p_0+c\,\tau>0$, die Approximation bleibt also von positivem Typ, wenn sie es vorher schon war. Aus diesem Grunde genügt es, sich auf

6 Bezüglich des Falles einer Raumdimension ($n=1$) vgl. man [4], S. 107ff.

Gln. (FP) zu beschränken (von denen (FPS) einen Sonderfall darstellt). Man erhält durch auf der Hand liegende Modifikationen der darzustellenden Theorie auch Aussagen über Approximationen nichtnegativen Typs für allgemeine inhomogene lineare parabolische Differentialgleichungen (8′).

Wir präzisieren den im Gitter $D = D(h, \tau)$ zu verwendenden Differenzenstern, d.h., das zu einem Punkt x im $\mathbb{R}^n$ gehörende System von Nachbargitterpunkten[7] $\mathfrak{N}(x) = \mathfrak{N}(x; h)$. Der Einfachheit halber sei $\{\mathfrak{N}(x)\}$ symmetrisch und translationsinvariant, also

(11a) $$\text{mit} \quad x \in \mathfrak{N}(y) \quad \text{sei} \quad y \in \mathfrak{N}(x),$$

(11b) $$\mathfrak{N}(x) = \mathfrak{N}(0) + x.$$

Der Zusammenhang mit dem Indexsystem $\mathfrak{R}$ ist gegeben durch

$$\mathfrak{R} = \{r \mid x + rh \in \mathfrak{N}(x, h)\}.$$

Statt (10) schreiben wir, mit der Bezeichnungsweise[7]

$$z = x + rh, \quad p_r(x, t; h, \tau) = p(z, x; t) = p(z, x; t; h, \tau)$$

(12) $$U(x, t+\tau) = \sum_{z \in \mathfrak{N}(x)} p(z, x; t)\, U(z, t), \quad t = \nu\tau,$$
$$\nu = 0, 1, 2, \ldots, [T/\tau] - 1.$$

Die Nichtnegativitätsbedingung (NN) geht über in

(13) $$p(x, y; t) \geqq 0 \quad \text{für} \quad y \in \mathfrak{N}(x).$$

Aus formalen Gründen ist es zweckmäßig, $p(x, y; t) = 0$ zu setzen für $y \notin \mathfrak{N}(x)$. Man kann dann in (12) und (14) unterm Summenzeichen $\mathfrak{N}(x)$ durch $\mathbb{R}^n \cap D$ ersetzen.

Denkt man sich $h^n U(z, t)$ als eine zur Zeit t im Gitterpunkt z befindliche Substanzmenge und $p(z, x; t)$ als den Bruchteil dieser Substanzmenge, der im Zeitintervall $(t, t+\tau)$ von z nach x wandert, so ist Substanzerhaltung gleichbedeutend mit der Bedingung

(14) $$\sum_{y \in \mathfrak{N}(x)} p(x, y; t) = 1.$$

Im Sinne dieser Auffassung liegt es, in der Anfangsbedingung

(15) $$U(x, 0) = G(x),$$

(16) $$G(x) = h^{-n} \int_{x - \varrho h}^{x + \varrho h} g(\xi)\, d\xi \quad \text{mit} \quad \varrho' = (1, 1, \ldots, 1)/2$$

zu nehmen. Dies ist $O(h^2)$-konsistent mit $u(x, 0) = g(x)$, da

$$|G(s) - g(s)| \leqq \left(\tfrac{1}{24} \sum_{j=1}^{n} R_{jj} + \tfrac{1}{16} \sum_{k=1}^{n} \sum_{j=1}^{k-1} R_{jk} \right) h^2$$

7 Aus Gründen einfacherer Schreibweise sei in Zukunft die Abhängigkeit des Nachbarschaftssystems und der Koeffizienten $p(z, x; t)$ von h und τ normalerweise nicht angeschrieben.

ist mit

$$R_{jk} = \max_{s-\varrho h \le x \le s+\varrho h} \left| \frac{\partial^2 g(x)}{\partial x_j \, \partial x_k} \right| .$$

Die Aufgabe ist nun die, (13) und (14) erfüllende konsistente Schemata zu finden. *Hierzu beschränken wir uns von jetzt ab auf einen einfachen Typ des Gitters*: $\mathfrak{N}(0) = \mathfrak{N}(0; h)$ *bestehe aus den* $1 + 2n^2$ *Punkten* 0, $\pm h\, e_j$, $\lambda h\, e_j + \mu h\, e_k$ *mit* $j \neq k$, *wobei* (λ, μ) *die Zahlenpaare* $(-1, -1)$, $(-1, 1)$, $(1, -1)$, $(1, 1)$ *durchläuft.* e_j *ist der Einheitsvektor in* x_j*-Richtung; zwecks Abkürzung der Schreibarbeit sei im folgenden* $h_j = h\, e_j$.

3. Diskretisierung der Fokker-Planck-Gleichung

Die unmittelbar sich anbietende, $O(h^2)$-konsistente Approximation von (FP) hat die Gestalt

$$\begin{aligned} &\frac{1}{\tau}\{U(x, t+\tau) - U(x, t)\} \\ &\quad = \frac{1}{2h^2} \sum_{j=1}^{n} \{b_{jj}(x + h_j, t)\, U(x + h_j, t) \\ &\quad\quad - 2 b_{jj}(x, t)\, U(x, t) + b_{jj}(x - h_j, t)\, U(x - h_j, t)\} \\ &\quad\quad + \frac{1}{4h^2} \sum_{j=1}^{n} \sum_{k=1}^{j-1} \sum_{\lambda, \mu} \lambda \mu\, b_{jk}(x + \lambda h_j + \mu h_k, t)\, U(x + \lambda h_j + \mu h_k, t) \\ &\quad\quad - \frac{1}{2h} \sum_{j=1}^{n} \{a_j(x + h_j, t)\, U(x + h_j, t) - a_j(x - h_j, t)\, U(x - h_j, t)\} . \end{aligned} \tag{17}$$

Hier und im folgenden bedeute $\sum_{\lambda, \mu}$ *stets Summation über die am Ende von Kapitel* 2 *angegebenen* 4 *Zahlenpaare* (λ, μ).

Der Diskretisierungsfehler ist gleich h^2 multipliziert mit einer gewichteten Summe aus vierten (auch gemischten) räumlichen Ableitungen der $b_{jk} u$, dritten räumlichen Ableitungen der $a_j u$ an Zwischenstellen (ξ, t), und der zweiten zeitlichen Ableitung von u an einer Zwischenstelle (x, s). Die Stellen ξ liegen in der konvexen Hülle von $\mathfrak{N}(x)$. Der Vollständigkeit halber sei hier erinnert an

$$\tau = 2\sigma h^2 . \tag{9}$$

Aus (9) und (17) ergibt sich

$$U(x, t+\tau) = \sum_{z \in \mathfrak{N}(x)} q(z, x; t)\, U(z, t) \tag{18}$$

mit

$$q(x, x; t) = 1 - 2\sigma \sum_{j=1}^{n} b_{jj}(x, t) , \tag{19a}$$

$$q(x, x \pm h_j; t) = \sigma\{b_{jj}(x, t) \pm h\, a_j(x, t)\} , \tag{19b}$$

$$q(x, x + \lambda h_j + \mu h_k; t) = \frac{\sigma}{2} \lambda \mu\, b_{jk}(x, t) , \quad j \neq k . \tag{19c}$$

Aus (17) findet man die Größen $q(z, x; t)$. Wegen der Symmetrie-Eigenschaft (11a) des Nachbarschaftssystems bereitet es aber keine Schwierigkeit, sich zu überlegen, wie die Größen $q(x, z; t)$ in (19) aussehen.

Die $q(x, y; t)$ erfüllen die Substanzerhaltungsbedingung; man rechnet leicht nach, daß

$$\sum_{y \in \mathfrak{N}(x)} q(x, y; t) = 1 \tag{14'}$$

ist. Die Nichtnegativitätsbedingung jedoch ist i. allg. leider nicht erfüllt, da unter den Größen (19c) negative auftreten, wenn nicht alle b_{jk} für $j \neq k$ verschwinden ($\lambda\mu$ nimmt ja die Werte $+1$ und -1 an). Um diesem Übelstand abzuhelfen, suchen wir Korrekturen $c(x, y; t)$ und Bedingungen ihrer Existenz derart, daß die Größen

$$p(x, y; t) = q(x, y; t) + c(x, y; t) \tag{20}$$

den Bedingungen (13) und (14) genügen. Es soll also gelten

$$q(x, y; t) + c(x, y; t) \geqq 0, \tag{13'}$$

$$\sum_{y \in \mathfrak{N}(x)} c(x, y; t) = 0. \tag{14'}$$

Ferner wünschen wir, daß dann (12) ebenso wie (18) zu (FP) $O(h^2)$-konsistent ist. Subtrahiert man auf beiden Seiten von (12) $U(x, t)$ und dividiert man dann durch τ, so ergibt sich wegen (9) und (20), daß für $O(h^2)$-Konsistenz die Bedingung

$$\sum_{z \in \mathfrak{N}(x)} c(z, x; t)\, v(z, t) = O(h^4), \qquad v(x, t) \text{ hinreichend glatt}, \tag{21}$$

notwendig und hinreichend ist.

Die Grundidee für die Anbringung der Korrekturen ist die, zuerst anstelle von (19c) nichtnegative Größen einzuführen. In (19b) und (19a) müssen dann die in (19c) angebrachten Korrekturen wieder kompensiert werden, und zwar gerade so, daß (21) gilt. Nichtnegativität der korrigierten Größen (19b) und (19a) erreicht man dann bei hinreichend starkem Überwiegen der Diagonalelemente $b_{jj}(x, t)$ über die Beträge der anderen Matrixelemente durch Wahl hinreichend kleiner h und σ.

Wir wählen für $j \neq k$ Funktionen $g_{jk}(x, t)$ mit der Symmetrieeigenschaft $g_{jk}(x, t) = g_{kj}(x, t)$, die zusammen mit ihren als existent vorausgesetzten Ableitungen nach x bis zur Ordnung 4 beschränkt seien. Dann erfüllen die Korrekturen

$$c(x, x; t) = \sigma \sum_{j=1}^{n} \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n} \right) g_{jk}(x, t), \tag{22a}$$

$$c(x, x \pm h_j; t) = -\sigma \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n} \right) g_{jk}(x, t), \tag{22b}$$

$$c(x, x + \lambda h_j + \mu h_k; t) = \frac{\sigma}{2} g_{jk}(x, t), \qquad j \neq k, \tag{22c}$$

die Bedingung (14') und, wie eine längere elementare Rechnung zeigt, auch die Bedingung (21). Die in (21) links stehende Summe ist gleich h^4 multipliziert mit einem Ausdruck, der Ableitungen der Funktionen $g_{jk}(x, t)$ und $u(x, t)$ nach x bis zur Ordnung 4 enthält, genommen an geeigneten Zwischenstellen der konvexen Hülle von $\mathfrak{N}(x)$.

Damit für $j \neq k$ die $p(x, x+\lambda h_j+\mu h_k; t) \geqq 0$ sind, fordern wir das Bestehen von

$$g_{jk}(x,t) \geqq |b_{jk}(x,t)|, \quad j \neq k. \tag{23}$$

Um $p(x, x \pm h_j; t) \geqq 0$ zu haben, muß, da i.allg. nicht alle a_j verschwinden, jedenfalls $b_{jj} > \sum_{k \neq j} g_{jk}$ sein. Man kommt zum Ziel, wenn man mit Konstanten $\delta > 0$, γ, $0 < \gamma < 1$, und einer Funktion $\eta(x,t) \geqq \delta$ das Erfülltsein der Bedingungen

$$|b_{jk}(x,t)| \leqq g_{jk}(x,t) \leqq |b_{jk}(x,t)| + \frac{\gamma\,\eta(x,t)}{n-1}, \quad j \neq k, \tag{24}$$

$$b_{jj}(x,t) \geqq \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n}\right) |b_{jk}(x,t)| + \eta(x,t), \quad j = 1, 2, \ldots, n, \tag{25}$$

fordert. Die Diagonaldominanz (25) ist die entscheidende einschränkende Bedingung für die Diffusionsmatrix $B(x,t)$. Aus ihr folgt nach dem Satz von Gerschgorin ([14], Kapitel 1), daß der kleinste Eigenwert von $B(x,t)$ stets $\geqq \delta$ ist.

Aus (24) und (25) folgt

$$p(x, x \pm h_j; t) \geqq \sigma\{(1-\gamma)\,\delta - h\,|a_j|\}$$

und dies ist $\geqq 0$, wenn

$$0 < h \leqq h_0 = (1-\gamma)\,\delta \Big/ \sup_{j,x,0 \leqq t \leqq T} |a_j(x,t)| \tag{26}$$

ist. Um $p(x, x; t) \geqq 0$ zu erhalten, nehme man

$$0 < \tau \leqq h^2 \Big/ \sup_{x,0 \leqq t \leqq T} \sum_{j=1}^{n} \left\{ b_{jj}(x,t) - \tfrac{1}{2}\left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n}\right) g_{jk}(x,t) \right\}. \tag{27}$$

Der Ausdruck hinter $\sup_{x,0 \leqq t \leqq T} \sum_{j=1}^{n}$ ist nach (24) und (25)

$$= \frac{1}{2} b_{jj} + \frac{1}{2}\left(b_{jj} - \sum_{k \neq j} g_{jk}\right) \geqq \frac{1}{2} b_{jj} + \frac{\eta}{4} > 0.$$

Mit Berücksichtigung von Satz 1 ergibt sich als Resultat

Satz 2. *Es existiere eine Funktion $\eta(x,t) \geqq \delta > 0$ so, daß für $j = 1, 2, \ldots, n$ die Diagonaldominanzbedingung* (25) *gilt. Man nehme für $j \neq k$ hinreichend glatte*[8] *Funktionen $g_{jk}(x,t)$, $g_{jk} = g_{kj}$, die den Bedingungen* (24) *genügen. Für h und τ sollen* (26) *und* (27) *erfüllt sein. Mit den Koeffizienten $p(x,y;t) = q(x,y;t) + c(x,y;t)$, q aus* (19a, b, c), *c aus* (22a, b, c), *ist dann* {(12), (15), (16)} *ein zu* {(FP), (A)} *$O(h^2)$-konsistentes und stabiles Differenzenschema, das nichtnegativitäts- und substanzerhaltend wirkt.*

Anmerkung 1. Eine Stabilitätskonstante C des Schemas (man vgl. das Stabilitätslemma 2 in Kapitel 2) findet man durch Betragssummenabschätzung aus der Stabilitätssumme $\sum_{z \in \mathfrak{N}(x)} p(z, x; t)$ (man vgl. Satz 1). Es ist ja

$$\sum_{z \in \mathfrak{N}(x)} p(z, x; t) \leqq 1 + C\,\tau. \tag{28}$$

8 Die g_{jk} und ihre Ableitungen nach x bis zur Ordnung 4 sollen existieren und beschränkt sein.

Eine längere Rechnung gibt

$$\sum_{z\in\mathfrak{N}(x)} p(z,x;t) = 1+\tau\left\{\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{\partial^2 \tilde{b}_{jk}}{\partial x_j\,\partial x_k} - \sum_{j=1}^{n}\frac{\partial \tilde{a}_j}{\partial x_j}\right.$$
$$\left. + \frac{h^2}{4}\sum_{j=1}^{n}\sum_{k=1}^{j-1}\left(\frac{\partial^4 \tilde{g}_{jk}}{\partial x_j^2\,\partial x_k^2} + \frac{1}{12}\frac{\partial^4 \tilde{g}_{jk}}{\partial x_k^4} - \frac{1}{6}\frac{\partial^4 \tilde{g}_{jk}}{\partial x_k^4} + \frac{1}{12}\frac{\partial^4 \tilde{g}_{jk}}{\partial x_k^4}\right)\right\}. \tag{29}$$

Anmerkung 2. Für den lokalen Abbruchfehler des Verfahrens (man vgl. Kapitel 2) ergibt sich nach längerer Rechnung

$$(L_D - L)\,u = \frac{\tau}{2}\frac{\partial^2 \tilde{u}}{\partial t^2} + h^2(B + A_3 + Q). \tag{30}$$

Mit $\gamma_{jk} = g_{jk}u$, $\alpha_j = a_j u$, $\beta_{jk} = b_{jk}u$ ist hier

$$B_4 = -\frac{1}{24}\sum_{j=1}^{n}\frac{\partial^4 \tilde{\beta}_{jj}}{\partial x_j^4} - \frac{1}{48}\sum_{j=1}^{n}\sum_{k=1}^{j-1}\left\{\left(\frac{\partial}{\partial x_j}+\frac{\partial}{\partial x_k}\right)^4\tilde{\beta}_{jk} - \left(\frac{\partial}{\partial x_j}-\frac{\partial}{\partial x_k}\right)^4\tilde{\beta}_{jk}\right\},$$

$$A_3 = \frac{1}{6}\sum_{j=1}^{n}\frac{\partial^3 \tilde{a}_j}{\partial x_j^3},$$

$$Q = \frac{1}{24}\sum_{j=1}^{n}\sum_{k=1}^{j-1}\left\{\frac{\partial^4 \tilde{\gamma}_{jk}}{\partial x_j^4} + \frac{\partial^4 \tilde{\gamma}_{jk}}{\partial x_k^4} - \frac{1}{2}\left(\frac{\partial}{\partial x_j}+\frac{\partial}{\partial x_k}\right)^4\tilde{\gamma}_{jk} - \frac{1}{2}\left(\frac{\partial}{\partial x_j}-\frac{\partial}{\partial x_k}\right)^4\tilde{\gamma}_{jk}\right\}.$$

Die Zeichen $\tilde{}$ bedeuten, daß die Ableitungen nach x an geeigneten Zwischenstellen (nicht jedesmal an denselben Stellen) der konvexen Hülle von $\mathfrak{N}(x)$ zu nehmen sind, die Ableitung nach t an einer Stelle (x, s) mit $t < s < t+\tau$.

Je nach Art der Anwendung des Mittelwertsatzes der Differentialrechnung kann man für Stabilitätssumme und Abbruchfehler natürlich auch andere Ausdrücke erhalten. Wesentlich sind die Ordnungen der auftretenden Ableitungen, aus denen man erkennen kann, was unter „hinreichend glatt" zu verstehen ist (man vgl. hierzu die in Kapitel 1 getroffene Vereinbarung).

Sonderfälle. Wenn für $j \neq k$ alle $b_{jk} \equiv 0$ sind, kann man $g_{jk} \equiv 0$ und $h_0 = \inf_{j,x,t}(b_{jj}/|a_j|)$ nehmen. Wenn für ein Paar (j,k) mit $j \neq k$ entweder stets $b_{jk} \geqq 0$ oder stets $b_{jk} \leqq 0$ ist, kann man das zugehörige $g_{jk} = |b_{jk}|$ nehmen (effektiv verkleinert sich dadurch der Differenzenstern um zwei Punkte). Wechselt keines der b_{jk} sein Vorzeichen und sind alle $a_j \equiv 0$, so kann man in (24) $\gamma = 1$, *in* (25) $\eta \equiv 0$ *zulassen und auf* (26) *verzichten (aus* (25) *folgt dann allerdings nur noch, daß B positiv semidefinit ist).*

Es sei noch erwähnt, daß bei Auffassung von (12) als Beschreibung eines diskreten Irrfahrtprozesses ($h^n U(x,t)$ = Wahrscheinlichkeit des Aufenthalts eines irrenden Teilchens am Orte x zur Zeit t, $p(x,z;t)$ = Wahrscheinlichkeit des Übergangs vom Orte x zum Orte z im Zeitintervall $(t, t+\tau)$, $h^n \sum_{x\in\mathbb{R}^n} G(x) = 1$) mit $\Delta t = \tau$ exakt die Relationen

$$\langle \Delta x\rangle = a(x,t)\,\Delta t, \qquad \langle \Delta x_j \Delta x_k\rangle = b_{jk}(x,t)\Delta t$$

gelten (man vgl. Kapitel 1).

4. Diskretisierung der Fokker-Planck-Gleichung mit selbstadjungiertem elliptischen Anteil

Die Differentialgleichung (FPS) ist ein Sonderfall von (FP), sie geht mit $a_j = \frac{1}{2}\sum_{k=1}^{n} \frac{\partial b_{jk}}{\partial x_k}$ in (FP) über. Man kann (FPS) also mit den Methoden des Kapitel 3 behandeln. Die spezielle Bauart von (FPS) empfiehlt aber ein anderes Vorgehen, bei dessen Darstellung wir uns aber kürzer fassen können.

Eine $O(h^2)$-konsistente Approximation kann man so gewinnen: Man formt (FPS) um in

$$\frac{\partial u}{\partial t} = \frac{1}{2}\sum_{j=1}^{n} \frac{\partial}{\partial x_j}\left(b_{jj}\frac{\partial u}{\partial x_j}\right) + \frac{1}{2}\sum_{j=1}^{n}\sum_{k=j+1}^{n}\left\{\frac{\partial}{\partial x_k}\left(b_{jk}\frac{\partial u}{\partial x_j}\right) + \frac{\partial}{\partial x_j}\left(b_{jk}\frac{\partial u}{\partial x_k}\right)\right\} \tag{31}$$

und verwendet die Beziehungen

$$\frac{\partial u}{\partial t} = \frac{1}{\tau}\{u(x, t+\tau) - u(x, t)\} + O(\tau),$$

$$\frac{\partial}{\partial x_j}\left(b_{jj}\frac{\partial u}{\partial x_j}\right) = \frac{1}{h^2}\left\{b_{jj}\left(x + \frac{1}{2}h_j, t\right)\left(u(x+h_j, t) - u(x, t)\right)\right.$$
$$\left. - b_{jj}\left(x - \frac{1}{2}h_j, t\right)\left(u(x, t) - u(x-h, t)\right)\right\} + O(h^2),$$

$$\frac{\partial}{\partial x_j}\left(b_{jk}\frac{\partial u}{\partial x_k}\right) = \frac{1}{4h^2}\{b_{jk}(x+h_j, t)\left(u(x+h_j+h_k, t) - u(x+h_j-h_k, t)\right)$$
$$- b_{jk}(x-h_j, t)\left(u(x-h_j+h_k, t) - u(x-h_j-h_k, t)\right)\} + O(h^2)$$

für $j \neq k$, und die Beziehung, die sich aus der letzten durch Vertauschung von j und k ergibt. Ersetzt man wie üblich u durch U, läßt die Terme $O(\tau)$ und $O(h^2)$ weg und löst dann nach $U(x, t+\tau)$ auf, so erhält man eine Beziehung der Gestalt (18) mit

$$q(x, x; t) = 1 - \sigma\sum_{j=1}^{n}\{b_{jj}(x + \tfrac{1}{2}h_j, t) + b_{jj}(x - \tfrac{1}{2}h_j, t)\}, \tag{32a}$$

$$q(x, x \pm h_j; t) = \sigma b_{jj}(x \pm \tfrac{1}{2}h_j, t), \tag{32b}$$

$$q(x, x + \lambda h_j + \mu h_k; t) = \frac{\sigma}{4}\lambda\mu\{b_{jk}(x + \lambda h_j, t) + b_{jk}(x + \mu h_k, t)\}, \quad j \neq k. \tag{32c}$$

Die Substanzerhaltungsbedingung $\sum_{y \in \mathfrak{N}(x)} q(x, y; t) = 1$ ist erfüllt, im allgemeinen aber wieder nicht die Nichtnegativitätserhaltungsbedingung $q(x, y; t) \geqq 0$. Eine Besonderheit des Schemas ist die Symmetrie

$$q(x, y; t) = q(y, x; t), \tag{33}$$

die ein diskretes Analogon der Selbstadjungiertheit des elliptischen Anteils ist.

Für die anzubringenden Korrekturen muß nun gelten

$$p(x, y; t) = q(x, y; t) + c(x, y; t) \geqq 0, \tag{34}$$

$$\sum_{y \in \mathfrak{N}(x)} c(x, y; t) = 0, \tag{35}$$

ferner eine Konsistenzbedingung, die wieder die Gestalt (21) hat, wenn das resultierende Schema $O(h^2)$-konsistent sein soll. Darüber hinaus wünschen wir Symmetrie

$$p(x, y; t) = p(y, x; t), \quad \text{also} \quad c(x, y; t) = c(y, x; t). \tag{36}$$

Man erreicht die gewünschten Ziele, wenn man wieder an die Matrix B eine geeignete Diagonaldominanzbedingung stellt und wieder hinreichend glatte Funktionen $g_{jk}(x, t)$ mit $g_{jk} = g_{kj}$ einführt, die der Bedingung (23) genügen. Man nehme

$$c(x, x; t) = \sigma \sum_{j=1}^{n} \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n} \right) g_{jk}(x, t), \tag{37a}$$

$$c(x, x \pm h_j; t) = -\frac{\sigma}{2} \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n} \right) \{ g_{jk}(x, t) + g_{jk}(x \pm h_j, t) \}, \tag{37b}$$

$$c(x, x + \lambda h_j + \mu h_k; t) = \frac{\sigma}{4} \{ g_{jk}(x + \lambda h_j, t) + g_{jk}(x + \mu h_k, t) \}, \quad j \neq k. \tag{37c}$$

Dann sind (36), (35) und (21) erfüllt. (34) ist erfüllt für die $p(x, x + \lambda h_j + \mu h_k; t)$, $j \neq k$. Damit (34) für die $p(x, x \pm h_j; t)$ gilt, muß für alle x des Gitters und alle j und für $\varkappa = \pm 1$

$$\frac{1}{2} \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n} \right) \{ g_{jk}(x, t) + g_{jk}(x + \varkappa h_j, t) \} \leqq b_{jj}\left(x + \frac{\varkappa}{2} h_j, t \right) \tag{38}$$

sein. Aus Symmetriegründen genügt es, (38) für $\varkappa = +1$ zu fordern. Addiert man die beiden Gln. (38) für $\varkappa = +1$ und $\varkappa = -1$ und beachtet, daß $g_{jk} \geqq 0$ ist, so erhält man

$$\left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n} \right) g_{jk}(x, t) \leqq b_{jj}(x + \tfrac{1}{2} h_j, t) + b_{jj}(x - \tfrac{1}{2} h_j, t). \tag{39}$$

Damit nun auch $p(x, x; t) \geqq 0$ ist, braucht man nur noch $\sigma = \tau/(2h^2)$ hinreichend klein zu wählen.

(38) ist erfüllbar, wenn mit einer Funktion $\eta(x, t) \geqq 0$ die Diagonaldominanzbedingung

$$b_{jj}(x + \tfrac{1}{2} h_j, t) \geqq \tfrac{1}{2} \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n} \right) \{ |b_{jk}(x, t)| + |b_{jk}(x + h_j, t)| \} + \eta(x, t), \tag{40}$$

$$j = 1, 2, \ldots, n,$$

erfüllt ist. Man nehme dann die g_{jk} so, daß gilt

$$|b_{jk}(x, t)| \leqq g_{jk}(x, t) \leqq |b_{jk}(x, t)| + \frac{\eta(x, t)}{n-1}, \quad j \neq k. \tag{41}$$

Als Resultat formulieren wir

Satz 3. *Es existiere eine Funktion $\eta(x, t) \geqq 0$ so, daß die Diagonaldominanzbedingung* (40) *gilt. Man nehme für $j \neq k$ hinreichend glatte Funktionen $g_{jk}(x, t)$, die den Bedingungen* (41) *und der Symmetriebedingung $g_{jk} = g_{kj}$ genügen. Es sei*

$h>0$ *und*

$$0<\tau\leqq 2h^2\Big/\sup_{x,0\leqq t\leqq T}\sum_{j=1}^{n}\Big\{b_{jj}(x+\tfrac{1}{2}h_j,t)+b_{jj}(x-\tfrac{1}{2}h_j,t)-\Big(\sum_{k=1}^{j-1}+\sum_{k=j+1}^{n}\Big)g_{jk}(x,t)\Big\}. \tag{42}$$

Mit den Koeffizienten $p(x,y;t)=q(x,y;t)+c(x,y;t)$, q *aus* (32a—c), c *aus* (37a—c), *ist dann* {(12), (15), (16)} *ein zu* {(FPS), (A)} $O(h^2)$*-konsistentes und stabiles Differenzenschema, das nichtnegativitäts- und substanzerhaltend wirkt.*

Anmerkung. Das Schema hat die Stabilitätskonstante $C=0$ (man vgl. Kapitel 2). Wegen (35), (36) folgt nämlich aus der Substanzerhaltung, daß die Stabilitätssumme $\sum_{z\in\mathfrak{N}(x)}p(z,x;t)\equiv 1$ ist. Der lokale Abbruchfehler $(L_D-L)u$ ist h^2 multipliziert mit einem Ausdruck, der von den Funktionen u, b_{jk}, g_{jk} Ableitungen nach x bis zur Ordnung 4, von u die zweite Ableitung nach t enthält. Diese Ableitungen sind an geeigneten Zwischenstellen genommen. Wir setzen sie alle als existent und beschränkt voraus, im Sinne der in der Einleitung getroffenen Vereinbarung über „hinreichend glatt". Wir verzichten, den ausgerechneten lokalen Abbruchfehler hier wiederzugeben.

Der Vollständigkeit halber sei noch erwähnt, daß bei Auffassung von (12) als Beschreibung eines diskreten Irrfahrtprozesses mit $\Delta t=\tau$ folgende Beziehungen gelten:

$$\begin{aligned}\langle\Delta x_j\rangle/\Delta t&=\frac{1}{2h}\Big\{b_{jj}\Big(x+\frac{1}{2}h_j,t\Big)-b_{jj}\Big(x-\frac{1}{2}h_j,t\Big)\Big\}\\&\quad+\frac{1}{4h}\Big(\sum_{k=1}^{j-1}+\sum_{k=j+1}^{n}\Big)\{b_{jk}(x+h_k,t)-b_{jk}(x-h_k,t)\}\\&=\frac{1}{2}\sum_{j=1}^{n}\frac{\partial b_{jk}(x,t)}{\partial x_k}+O(h^2),\end{aligned}$$

$$\begin{aligned}\langle\Delta x_j\Delta x_j\rangle/\Delta t&=\frac{1}{2}\Big\{b_{jj}\Big(x+\frac{1}{2}h_j,t\Big)+b_{jj}\Big(x-\frac{1}{2}h_j,t\Big)\Big\}-\frac{1}{4}\Big(\sum_{k=1}^{j-1}+\sum_{k=j+1}^{n}\Big)\\&\quad\{g_{jk}(x+h_k,t)-2g_{jk}(x,t)+g_{jk}(x-h_k,t)\}\\&=b_{jj}(x)+O(h^2),\end{aligned}$$

$$\begin{aligned}\langle\Delta x_j\Delta x_k\rangle/\Delta t&=\frac{1}{4}\{b_{jk}(x+h_j,t)+b_{jk}(x-h_j,t)+b_{jk}(x+h_k,t)+b_{jk}(x-h_k,t)\}\\&=b_{jk}(x,t)+O(h^2),\qquad j\neq k.\end{aligned}$$

5. Wie notwendig sind die Diagonaldominanzbedingungen?

Wir fragen, ob für beliebig kleines h Differenzenapproximationen nichtnegativen Typs mit dem Nachbarschaftssystem $\mathfrak{N}(x;h)$ für die parabolische Differentialgleichung

$$\frac{\partial u}{\partial t}=\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n}b_{jk}(x,t)\frac{\partial^2 u}{\partial x_j\,\partial x_k}+\sum_{j=1}^{n}\tilde{a}_j(x,t)\frac{\partial u}{\partial x_j}+\tilde{c}(x,t)\,u \tag{8}$$

existieren können, wenn die abgeschwächte Diagonaldominanzbedingung

$$b_{jj}(x,t) \geqq \left(\sum_{k=1}^{j-1} + \sum_{k=j+1}^{n}\right) |b_{jk}(x,t)|, \quad j = 1, 2, \ldots, n, \tag{43}$$

in einer einen Punkt (ξ, s) enthaltenden offenen Menge des $\mathbb{R}^n \times \{t \mid 0 \leqq t \leqq T\}$ verletzt ist, d.h., wenn dort für ein $j = j_0$ in (43) statt $\geqq$ das Zeichen $<$ steht. Da man die Koordinaten x_j umnumerieren kann, können wir ohne Beschränkung der Allgemeinheit $j_0 = 1$ annehmen. Es genügt, diese Frage für (8) zu untersuchen, da sowohl (FP) als auch (FPS) in diese Gestalt umgeformt werden können. Ferner ist es geboten, das in (25) und (40) auftretende δ durch 0 zu ersetzen, da in Sonderfällen $\delta = 0$ erlaubt ist (man vgl. den Abschnitt „Sonderfälle" in Kapitel 3).

Wegen der großen Anzahl der Konsistenzgleichungen, zu denen die Nichtnegativitätsbedingungen noch hinzukommen, scheint das Problem für $n \geqq 3$ sehr schwierig zu sein. Der Fall $n = 2$ läßt sich behandeln durch Modifikation einer von Greenspan für elliptische Differentialgleichungen entworfenen Methode ([8], S. 62—70). Es gilt der

Satz 4. *Die Differentialgleichung* (8) *mit* $n = 2$ *sei gleichmäßig parabolisch. Die* $b_{jk} = b_{kj}$, $\tilde{a}_j$ *und* $\tilde{c}$ *seien hinreichend glatt*[9], *speziell seien sie beschränkt. In einer einem Punkt* (ξ, s) *enthaltenden offenen Menge* M *sei*

$$b_{11}(x,t) < |b_{12}(x,t)|. \tag{44}$$

Dann existiert in diesem Punkt für genügend kleines h *keine Differenzenapproximation nichtnegativen Typs zu* (8) *mit dem Nachbarschaftssystem* $\mathfrak{N}(x;h) = \{x + \varrho_1 h_1 + \varrho_2 h_2 \mid \varrho_1 = -1, 0, 1;\ \varrho_2 = -1, 0, 1\}$.

Beweis. Für genügend kleines h können wir annehmen, daß die konvexe Hülle von $\mathfrak{N}(\xi, h) \times \{t \mid s - \tau \leqq t \leqq s + \tau\}$ in M liegt. Wir greifen den Punkt (ξ, s) aus M heraus, machen $x = \xi$ zum Zentrum einer Familie $\mathfrak{N}(x; h)$ von Nachbarschaftssystemen und beachten die Konstanz von σ in (9). Eine konsistente Approximation (10) nichtnegativen Typs muß die Bedingungen (K1), (K2), (K3), (NN) erfüllen mit $\mathfrak{N} = \{r \mid x + rh \in \mathfrak{N}(x; h)\}$.

Wegen (K1) und (NN) sind die p_r beschränkt, für eine geeignete Nullfolge $\{h_\varkappa\}$, $\varkappa = 1, 2, \ldots, \to \infty$, streben also die $p_r(x, s; h_\varkappa, \tau)$ gegen Grenzwerte $\beta_r = \beta_r(x, s)$. Dabei gehen (K1), (K2), (K3) über in

$$\sum_{r \in \mathfrak{N}} \beta_r = 1, \quad \sum_{r \in \mathfrak{N}} r_j \beta_r = 0, \quad \sum_{r \in \mathfrak{N}} r_j r_k \beta_r = 2\sigma b_{jk}. \tag{45}$$

(NN) geht über in

$$\beta_r \geqq 0 \quad \text{für} \quad r \in \mathfrak{N}. \tag{46}$$

Setzt man

$$\alpha_0 = \beta_{\binom{0}{0}} - 1, \quad \alpha_m = \beta_{r'} \quad \text{für} \quad r' \neq (0, 0), \tag{47}$$

wobei der einfache Index m der Reihe nach die Werte 1, 2, 3, 4, 5, 6, 7, 8 annimmt, wenn der Doppelindex r' der Reihe nach die Paare (1, 0), (0, 1), (−1, 0), (0, −1), (1, 1), (−1, 1), (−1, −1), (1, −1) annimmt, so geht (45) über in das

9 So glatt, wie es für die Erfordernisse der Konsistenztheorie erforderlich ist.

System

$$
\begin{aligned}
&+\alpha_0+\alpha_1+\alpha_2+\alpha_3+\alpha_4+\alpha_5+\alpha_6+\alpha_7+\alpha_8=0\\
&+\alpha_1\qquad-\alpha_3\qquad+\alpha_5-\alpha_6-\alpha_7+\alpha_8=0\\
&\qquad+\alpha_2\qquad-\alpha_4+\alpha_5+\alpha_6-\alpha_7-\alpha_8=0\\
&+\alpha_1\qquad+\alpha_3\qquad+\alpha_5+\alpha_6+\alpha_7+\alpha_8=2\sigma b_{11}\\
&\qquad\qquad\qquad\qquad+\alpha_5-\alpha_6+\alpha_7-\alpha_8=2\sigma b_{12}\\
&\qquad+\alpha_2\qquad+\alpha_4+\alpha_5+\alpha_6+\alpha_7+\alpha_8=2\sigma b_{22},
\end{aligned}
\tag{48}
$$

das mit α_6, α_7, α_8 als Parametern eindeutig nach den Größen α_0, α_1, α_2, α_3, α_4, α_5 aufgelöst werden kann. Wir verzichten darauf, die vollständige Lösung anzuschreiben, sondern begnügen uns mit

$$\alpha_1=\sigma b_{11}-2\sigma b_{12}-\alpha_6+\alpha_7-2\alpha_8 \tag{49}$$

und

$$\alpha_3=\sigma b_{11}-\alpha_6-\alpha_7. \tag{50}$$

(49), (50) und die Ungleichungen

$$\alpha_0\leqq 0,\quad \alpha_m\geqq 0 \quad \text{für} \quad m=1,2,\ldots,8 \tag{51}$$

müssen erfüllt sein, wenn die Approximation von nichtnegativem Typ sein soll, was wir als Antithese zur Behauptung des Satzes 4 annehmen wollen.

Ohne Beschränkung der Allgemeinheit können wir nun voraussetzen, daß $b_{12}(x,s)>0$ ist. Denn wenn $b_{12}<0$ ist, kann man (8) durch die Variablensubstitution $\xi_1=-x_1$, $\xi_2=x_2$ in eine parabolische Differentialgleichung gleicher Struktur mit β_{jk} anstelle von b_{jk} transformieren, für die aber im interessierenden Punkt $\beta_{12}>0$ ist.

Wir nehmen also an, es sei $b_{12}>b_{11}$ im Punkte (x,s). Aus (51) und (49) folgt dann $2\alpha_8-\alpha_7+\alpha_6\leqq\sigma b_{11}-2\sigma b_{12}<-\sigma b_{11}$, also $\alpha_7>\sigma b_{11}+2\alpha_8+\alpha_6$, mithin wegen (51)

$$\alpha_7>\sigma b_{11}. \tag{52}$$

Andererseits ist wegen (51) und (50) $\alpha_7\leqq\sigma b_{11}-\alpha_6\leqq\sigma b_{11}$, also im Widerspruch zu (52)

$$\alpha_7\leqq\sigma b_{11}. \tag{53}$$

Damit ist der Satz bewiesen.

6. Beispiele

Die Sätze 2 und 3 erfordern für $j\neq k$ die Angabe von Korrekturfunktionen g_{jk}. Dies ist sehr einfach, falls keine der Funktionen b_{jk} ihr Vorzeichen wechselt. Falls die Funktion b_{jk} ihr Vorzeichen wechselt, bestimme man g_{jk} in der Umgebung des Nulldurchgangs von b_{jk} durch genügend glatte Abrundung der „Kante" des Graphen der Funktion $|b_{jk}|$, dabei ist auf die Bedingung (24) bzw. (41) zu achten. Falls die Diagonaldominanz der Matrix $B=(b_{jk})$ nur schwach ausgeprägt ist, kann dies eine unangenehme Aufgabe sein.

Zur Illustration diene die Matrix (mit $n=3$)

$$B(x,t)=\begin{pmatrix} 5 & \sin(x_1+x_3) & \cos t \\ \sin(x_1+x_3) & \frac{7}{2} & e^{-t} \\ \cos t & e^{-t} & 4 \end{pmatrix}.$$

Aus (25) folgt

$$\begin{aligned}\delta &\leqq \eta(x,t)\\ &\leqq \min\{5-|\sin(x_1+x_3)|-|\cos t|,\ \tfrac{7}{2}-|\sin(x_1+x_3)|-e^{-t},\ 4-|\cos t|-e^{-t}\}.\end{aligned}$$

Eine mögliche Wahl ist $\eta(x,t)\equiv\frac{3}{2}=\delta$. Diese Wahl genügt auch der Bedingung (40). Für die g_{jk} betrachten wir nur die stärkere Bedingung (24), die hier so lautet

$$|b_{jk}(x,t)|\leqq g_{jk}(x,t)\leqq|b_{jk}(x,t)+\tfrac{1}{2},\qquad \gamma=\tfrac{2}{3}.$$

Kritisch sind die Nulldurchgänge von $b_{12}=\sin(x_1+x_3)$ und $b_{13}=\cos t$. Man kommt zum Ziel mit

$$g_{12}(x,t)=\tfrac{1}{2}+\sin^2(x_1+x_3),\qquad g_{23}(x,t)=e^{-t},\qquad g_{13}=\tfrac{1}{2}+\cos^2 t.$$

Mit $s=x_1+x_3$ ist nämlich

$$\begin{aligned}g_{12}-|b_{12}| &= \tfrac{1}{2}+\sin^2 s-|\sin s|=\tfrac{1}{2}-|\sin s|\,(1-|\sin s|),\\ g_{23}-|b_{23}| &= \tfrac{1}{2}+\cos^2 s-|\cos s|=\tfrac{1}{2}-|\cos s|\,(1-|\cos s|)\end{aligned}$$

und für $0\leqq y\leqq 1$ ist $0\leqq y(1-y)\leqq\frac{1}{4}$, mit $y=|\sin s|$ folgt also

$$\tfrac{1}{4}\leqq g_{12}-|b_{12}|\leqq\tfrac{1}{2},\qquad |b_{12}|+\tfrac{1}{4}\leqq g_{12}\leqq|b_{12}|+\tfrac{1}{2},$$

(24) ist also erfüllt für g_{12}. Analog zeigt man, daß (24) für g_{13} erfüllt ist.

Abschließend sei für den Fall $n=2$ ein numerisches Beispiel betrachtet, und zwar eine inhomogene Gleichung, deren Lösung bekannt ist:

$$\frac{\partial u}{\partial t}=\frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{\partial}{\partial x_j}\left(b_{jk}\frac{\partial u}{\partial x_k}\right)+f$$

mit

$$B=\begin{pmatrix} \frac{3}{2} & \sin(x_1+x_2) \\ \sin(x_1+x_2) & 2 \end{pmatrix},$$

$$f=2t\{2+\sin(x_1+x_2)\}+(\tfrac{1}{4})(1+t^2)\{7\sin(x_1+x_2)-4\cos(2(x_1+x_2))\}$$

und der Lösung

$$u(x,t)=(1+t^2)\{2+\sin(x_1+x_2)\}.$$

Die Gleichung hat selbstadjungierten elliptischen Anteil, und wir diskretisieren gemäß Kapitel 4 mit der Modifikation (10′) (mit $F=f$). Wir nehmen $\eta(x,t)=\frac{3}{2}-|\sin(x_1+x_2)|$ und, verträglich mit (41), $g_{12}(x,t)=1$. Damit finden wir mit den Formeln (32), (34), (37)

$$\begin{aligned}&p(x,x;t)=1-5\sigma,\qquad p(x,x\pm h_1;t)=\sigma/2,\qquad p(x,x\pm h_2;t)=\sigma,\\ &p(x,x+\lambda h_1+\mu h_2;t)=(\sigma/4)\{\lambda\mu(\sin(x_1+x_2+\lambda h)+\sin(x_1+x_2+\mu h))+2\},\end{aligned}$$

und haben, mit $\sigma \leqq \frac{1}{6}$ und $x = mh$,

$$U(x, 0) = g(x) = 2 + \sin(x_1 + x_2),$$
$$U(x, t+\tau) = \sum_{z \in \mathfrak{N}(x)} p(z, x; t)\, U(z, t) + \tau f(x, t).$$

Bei der Rechnung beachten wir die Periodizität der Diffusionsmatrix $B(x, t)$ und der Anfangsbedingung und beschränken uns auf das Quadrat $-\pi \leqq x_1 \leqq \pi$, $-\pi \leqq x_2 \leqq \pi$ mit periodischer Fortsetzung der $p(z, x; t)$ und der jeweils erhaltenen Werte $U(x, t)$.

Die Rechnung[10] wurde durchgeführt auf der IBM 360/91 des Instituts für Plasmaphysik mit ungefähr 16 signifikanten Dezimalstellen, und zwar mit den Schrittweiten $h = \pi/16$ und $h = \pi/32$ und dem Parameter $\sigma = 0{,}15$. Für die Werte $t = 2$ und $t = 4$ und die Stellen $x = (j_1\pi/4,\ j_2\pi/4)$, j_1, j_2 ganzzahlig, wurden die Werte $U - u$ für beide Schrittweiten ermittelt und durcheinander dividiert. Da der Fehler asymptotisch wie $O(h^2)$ abnimmt, müssen diese Quotienten in der Nähe von 4 liegen. Infolge der Symmetrie-Eigenschaften der speziell gewählten Differentialgleichung und ihrer Lösung treten in jedem Resultat-Schema nur fünf voneinander verschiedene Einträge auf, die beispielsweise an den Stellen $x = (0,\ j_2\pi/4)$ stehen, wobei $j_2 = -2, -1, 0, 1, 2$. Die gerundeten Werte geben wir in Form einer Tabelle:

t	j_2	U mit $h = \pi/16$	U mit $h = \pi/32$	u	Fehlerquotient
2	−2	5,01079	5,00269	5,00000	4,01
2	−1	6,46350	6,46427	6,46447	4,84
2	0	10,02634	10,00668	10,00000	3,94
2	1	13,62725	13,55841	13,53553	4,01
2	2	15,12501	15,03109	15,00000	4,02
4	−2	17,00923	17,00226	17,00000	4,08
4	−1	21,93758	21,96889	21,97918	4,04
4	0	34,02561	34,00671	34,00000	3,81
4	1	46,24131	46,07578	46,02082	4,01
4	2	51,32175	51,07997	51,00000	4,02

Der in der zweiten Zeile auftretende ungewöhnlich große Fehlerquotient 4,84 hat offenbar die Ursache, daß $U - u$ dort ein anderes Vorzeichen hat als an den anderen Stellen (x, t), für die Resultate angegeben sind.

Zusammenfassung

Für die partielle Differentialgleichung von Fokker und Planck im $\mathbb{R}^n$ (Vorwärts-Differentialgleichung von Kolmogorov) mit orts- und zeitabhängiger Drift und Diffusionsmatrix werden, unter einigen zusätzlichen Bedingungen, explizite Differenzenschemata angegeben, die zwei wesentliche Eigenschaften des beschriebenen Diffusionsprozesses imitieren, nämlich Nichtnegativitätserhaltung und Substanzerhaltung. Diese Schemata können auch interpretiert werden als Beschreibung diskreter instationärer inhomogener Irrfahrten. Sie sind stabil in der

10 Für die Programmierung bin ich Herrn J. Steuerwald zu Dank verpflichtet.

Maximumnorm, und man kann mit ihrer Hilfe bei gegebener Anfangsbedingung die Differentialgleichung numerisch approximativ lösen oder aber den durch sie beschriebenen Diffusionsprozeß approximativ simulieren (Monte-Carlo) und veranschaulichen. Wesentliche Bedingungen sind die Beschränktheit der Koeffizienten der Differentialgleichung und einige ihrer Ableitungen und eine starke Diagonaldominanz der Diffusionsmatrix. Die Resultate lassen sich verallgemeinern auf allgemeine lineare parabolische Differentialgleichungen (Substanzerhaltung ist dann i.allg. nicht mehr vorhanden). Die beschriebene Methode erlaubt auch, für lineare elliptische Operatoren im $\mathbb{R}^n$ Approximationen nichtnegativen Typs anzugeben, wenn diese analogen Bedingungen genügen.

Literatur

1. Amann, H.: Monte-Carlo-Methoden und lineare Randwertprobleme. Zeitschrift für Angewandte Mathematik und Mechanik **48**, 109—116 (1968).
2. Bers, L.: On mildly non-linear partial differential equations of elliptic type. Jour. Res. Nat. Bur. Stand **51**, 229—236 (1953).
3. Bharucha-Reid, A. T.: Elements of the theory of Markov processes and their applications. New York-Toronto-London: McGraw Hill, Inc. 1960.
4. Forsythe, G. E., Wasow, W. R.: Finite difference methods for partial differential equations. New York-London: John Wiley & Sons, Inc. 1960.
5. Feller, W.: An Introduction to probability theory and its applications, volume 1, third edition. New York-London-Sydney: John Wiley & Sons, Inc. 1968.
6. Friedman, A.: Partial differential equations of parabolic type. Englewood Cliffs, N. J.: Prentice-Hall, Inc. 1964.
7. Gorenflo, R.: Differenzenschemata vom Irrfahrttypus für die Differentialgleichung von Fokker-Planck-Kolmogorov. Zeitschrift für Angewandte Mathematik und Mechanik **48**, T 69—T 72 (1968), Sonderheft GAMM-Tagung.
8. Greenspan, D.: Introductory numerical analysis of elliptic boundary value problems. New York-Evanston-London: Harper & Row 1965.
9. Henze, E.: Räumlich homogene Irrfahrten im Gitter, Teil II: Instationäre Irrfahrten. Mathematische Annalen **147**, 422—444 (1962).
10. Isaacson, E., Keller, H. B.: Analysis of numerical methods. New York-London-Sydney: John Wiley & Sons, Inc. 1966.
11. Kac, M.: Random walk and the theory of Brownian motion. American Mathematical Monthly **54**, 369—391 (1947).
12. Motzkin, T. S., Wasow, W.: On the approximation of linear elliptic differential equations by difference equations with positive coefficients. Journal of Mathematics and Physics **31**, 253—259 (1952/53).
13. Lord Rayleigh: On James Bernoulli's theorem in probabilities. Philosophical Magazine and Journal of Science **47**, 246—251 (1899).
14. Varga, R. S.: Matrix Iterative Analysis. Englewood Cliffs, N. J.: Prentice-Hall, Inc. 1965.

R. Gorenflo
Institut für Plasmaphysik GmbH
8046 Garching bei München

Numer. Math. 14, 468—475 (1970)

Eine Möglichkeit zur Konvergenzbeschleunigung bei Iterationsverfahren für bestimmte nichtlineare Probleme

Dietrich Braess

Eingegangen am 26. Februar 1969

Nichtlineare Gleichungssysteme, nichtlineare Approximationsaufgaben und andere nichtlineare Probleme behandelt man oft numerisch mit Iterationsverfahren. In jedem Iterationsschritt ist eine lineare Aufgabe zu lösen, die eine Näherung des vorliegenden Problems darstellt. Solche Verfahren lassen sich in einem allgemeinen Rahmen formulieren. Eine Möglichkeit zur Konvergenzbeschleunigung bietet sich dann bei Problemen, die in einem bestimmten Sinne teilweise linear sind, indem der klassische Algorithmus durch einen kleinen Optimierungsprozeß erweitert wird. Für Probleme mit Exponentialsummen erhält man diese Form der Beschleunigung auch mit Hilfe von Invarianzbetrachtungen.

1

Nichtlineare Probleme behandelt man oft mit Hilfe von Iterationsverfahren. Mit jedem Schritt wird eine lineare Aufgabe numerisch gelöst, die als Näherung des zugrunde liegenden Problems verstanden werden kann. Das bekannteste Verfahren von diesem Typ ist das Newtonsche Verfahren zur Lösung nichtlinearer Gleichungssysteme.

$$F_i(\alpha) \equiv F_i(\alpha_1, \alpha_2 \dots \alpha_n) = 0 \qquad (i = 1, 2 \dots n). \tag{1}$$

Die Gleichungen fassen wir durch Einführung der n-wertigen Funktion $F[\alpha]$ zusammen.

$$F[\alpha] = 0.$$

Zu der Näherung beim ν-ten Iterationsschritt α^ν wird die (differenzierbare) Funktion $F[\alpha]$ durch die lineare Funktion

$$\widetilde{F}[\alpha] = F[\alpha^\nu] + (\alpha - \alpha^\nu)' \nabla F[\alpha^\nu] \tag{2}$$

formal ersetzt, und als nächste Näherung $\alpha^{\nu+1}$ wird der Lösungsvektor des linearen Systems

$$\widetilde{F}_i(\alpha) = 0 \qquad (i = 1, 2 \dots n) \tag{3}$$

betrachtet. Nach dem gleichen Prinzip gelangt man zu Newtonschen Verfahren für andere Aufgaben, z.B. für die Anpassung nichtlinearer Funktionen nach der Methode der kleinsten Quadrate [1, 6]. Das Konstruktionsprinzip formulieren wir im nächsten Abschnitt so allgemein, daß sich u.a. auch die Remes-Algorithmen für die nichtlineare Tschebyscheff-Approximation [7, 10, 11] einordnen.

Bekanntlich braucht bei solchen Verfahren in vielen Fällen keine Konvergenz einzutreten, wenn die Ausgangsnäherung nicht schon hinreichend nahe bei der

Lösung gewählt ist. Eine Möglichkeit zur Konvergenzbeschleunigung bietet sich bei „teilweise linearen" Problemen, d.h. wenn die Funktion $F[\alpha]$ von einigen Parametern linear abhängt. So sind z.B. bei der Exponentialapproximation die Funktionen

$$F(x, \alpha) = \sum_{k=1}^{l} \alpha_k e^{\alpha_{k+l} x} \qquad (n = 2l) \tag{4}$$

linear in den ersten l Parametern. Bei der Anwendung der allgemeinen Vorschrift, die im dritten Abschnitt beschrieben wird, sind bei der modifizierten Iteration die linearen Parameter[1] durch einen Optimierungsprozeß zu bestimmen, und nur die übrigen Parameter[2] werden in der üblichen Weise berechnet. Im vierten Abschnitt folgen Invarianzbetrachtungen für die Iteration bei der Exponentialapproximation. Im fünften Abschnitt werden noch zwei weitere Beispiele für die Anwendung des Konstruktionsprinzips gegeben.

Der Erfolg der modifizierten Iteration zeigte sich gerade bei der Anpassung mit Exponentialsummen sehr deutlich, weil das übliche Newtonsche Verfahren bei mehrgliedrigen Summen wegen der Rundungsfehler meistens versagt [2].

2

Ehe wir die Verfahren allgemein formulieren, werden die Voraussetzungen an Hand eines Algorithmus für die nichtlineare Tschebyscheff-Approximation erläutert. Gegeben sei eine Menge von Funktionen $F[\alpha] = F(x, \alpha)$, die auf einem reellen Intervall X definiert und die durch den Parametervektor α charakterisiert sind. Zu einer vorgelegten Funktionen $f(x)$ ist für den Anpassungsfehler

$$\Phi(\alpha) = \|F[\alpha] - f\| := \max_{x \in X} |F(x, \alpha) - f(x)| \tag{5}$$

das Minimum zu ermitteln. Die Bestimmung des optimalen Parametervektors ist nun selbst dann nicht in einem Schritt möglich, wenn $F[\alpha]$ linear von den Parametern abhängt.

Eine starke Vereinfachung ergibt sich, wenn man die Anpassung nur über einer endlichen und diskreten Punktmenge $\{x_i\}$ und nicht über dem ganzen Intervall ausführt. Im linearen Fall ist die Minimierung von

$$\max_{x_i} |F(x_i, \alpha) - f(x_i)| \tag{6}$$

für endliche Punktmengen $\{x_i\} \subset X$ mit Hilfe eines linearen Programms möglich, das sich in Spezialfällen sogar auf ein lineares Gleichungssystem reduziert.

Bei der Konstruktion der besten Approximation nach Remes wählt man abhängig von der Näherung α^ν eine Menge $X^\nu \subset X$ mit (mindestens) $n+1$ Punkten, die im allgemeinen Extremalpunkte der Fehlerkurve $F(x, \alpha^\nu) - f(x)$ sind. Die Fehlerfunktion

$$\Phi^\nu(\alpha) = \max_{x \in X^\nu} |F(x, \alpha) - f(x)| \tag{7}$$

besitzt, wenn man sie auch als Funktional von F auffaßt, folgende Eigenschaften.

1 Zum Beispiel sind das in (4) die Parameter $\alpha_1 \dots \alpha_l$.

2 Zum Beispiel sind das in (4) die Parameter $\alpha_{l+1} \dots \alpha_{2l}$.

I. Für $\alpha = \alpha^\nu$ gilt

$$\Phi^\nu(\alpha^\nu) = \Phi(\alpha^\nu). \tag{8}$$

II. Zu jeder linearen Funktion

$$F[\alpha] = H_0 + \sum_{k=1}^{l} \alpha_k H_k$$

läßt sich numerisch das Minimum von $\Phi^\nu(\alpha)$ bestimmen. Dabei ist für die Anzahl der (freien) Parameter α_k jede Zahl $l \leqq n$ zulässig.

Unter diesen Voraussetzungen läßt sich die Iterationsvorschrift für Fréchet-differenzierbare Funktionen $F[\alpha]$ allgemein formulieren.

1. Für $\alpha = \alpha^\nu$ werden zum Aufbau der „linearisierten Funktion" $\tilde{F}$ die Ableitungen $\frac{\partial}{\partial \alpha_k} F[\alpha^\nu]$ ermittelt.

2. Wenn in dem Funktional Φ^ν an Stelle von F die Näherung $\tilde{F}$ eingesetzt wird, erhält man die Funktion $\tilde{\Phi}^\nu(\alpha)$. Für diese ist das Minimum $\alpha = \tilde{\alpha}$ zu ermitteln. Es sei

$$\Delta\alpha := \tilde{\alpha} - \alpha^\nu \tag{9a}$$

die berechnete Parameteränderung und

$$\Delta\Phi := \Phi(\alpha^\nu) - \tilde{\Phi}^\nu(\tilde{\alpha}) = \tilde{\Phi}^\nu(\alpha^\nu) - \tilde{\Phi}^\nu(\tilde{\alpha}) \tag{9b}$$

die für das linearisierte Problem erreichte Verbesserung.

3. Es werde zunächst der Dämpfungsfaktor $t = 1$ gesetzt und getestet, ob

$$\Phi(\alpha^\nu + t\Delta\alpha) \leqq \Phi(\alpha^\nu) - \tfrac{1}{3} t \Delta\Phi \tag{10}$$

gilt. Sofern die Bedingung (10) erfüllt ist, wird ein neuer Iterationsschritt mit

$$\alpha^{\nu+1} = \alpha^\nu + t\Delta\alpha \tag{11}$$

begonnen. Andernfalls wird der Dämpfungsfaktor t halbiert und der Teilschritt 3 wiederholt[3].

Die Iteration erzeugt eine monotone Folge

$$\Phi(\alpha^0) \geqq \Phi(\alpha^1) \geqq \Phi(\alpha^2) \geqq \cdots. \tag{12}$$

Sie wird beendet[3], sofern wegen der Rundungsfehler keine Verbesserung mehr zu erreichen ist.

Bei den klassischen (Newtonschen) Iterationsverfahren gibt es nur die ersten beiden Teilschritte, und es wird direkt $\alpha^{\nu+1} = \tilde{\alpha} = \alpha^\nu + \Delta\alpha$ gesetzt. Das entspricht einem Dämpfungsfaktor $t = 1$[4].

Die Monotonie (12) wird durch den Test im dritten Teilschritt erreicht. An die Stelle des Tests tritt bei einigen Autoren die Minimierung von $\Phi(\alpha^\nu + t\Delta\alpha)$ zur Bestimmung des Faktors t [4, 6]. Bei einer anderen Variante wird der Faktor

3 Daß in Programmen für Rechenmaschinen vorsichtshalber für die Zahl der Iterationsschritte und Teilschrittwiederholungen eine Schranke einzubauen ist, lassen wir in dem formalen Programm außer acht.

4 Die im folgenden dargestellte Modifikation zur Konvergenzbeschleunigung kann auch ohne Dämpfung (d.h. für $t = 1$) verwandt werden, wenn auch dort der Test mit seinen Konsequenzen unterdrückt wird.

mit Hilfe der zweiten Ableitungen so abgeschätzt, daß eine zu (10) äquivalente Relation erfüllt ist [8]. In jedem Fall ist die Iteration in elektronischen Rechnern wegen (12) leicht zu kontrollieren, da im Gegensatz zu den klassischen Verfahren die bekannte Käfigbildung und andere Oscillationen [9] nicht auftreten. Die Konvergenz erfolgt in einer Form, wie sie von Ostrowski für Gradientenverfahren beschrieben wurde [8].

Eine Gefahr dafür, daß die Iteration vorzeitig abgebrochen wird, besteht allerdings bei kleinen Dämpfungsparametern. Denn $\Phi(\alpha^\nu)$ und $\Phi(\alpha^\nu + t\,\Delta\alpha)$ unterscheiden sich dann nur wenig, und die gerundeten numerischen Werte erfüllen nicht mehr die Relation (10). Aus diesem Grunde erwies sich eine Modifikation zur Konvergenzbeschleunigung als zweckmäßig.

3

Das im letzten Abschnitt beschriebene Konstruktionsprinzip modifizieren wir bei Problemen, bei denen die Funktion $F[\alpha]$ von einigen Parametern α_k linear abhängt, also in der Form (13) darstellbar ist

$$F[\alpha] = \sum_{k=1}^{l} \alpha_k H_k[\alpha_{l+1}, \alpha_{l+2} \dots \alpha_n] \qquad (l < n). \tag{13}$$

Um diese Struktur noch deutlicher hervorzuheben, teilen wir die Parameter in zwei Gruppen und schreiben mit $\alpha = (\beta, \gamma)$

$$F[\alpha] = F[\beta, \gamma] = \sum_{k=1}^{l} \beta_k H_k[\gamma]$$

und entsprechend

$$\Phi(\alpha) = \Phi(\beta, \gamma), \qquad \Phi^\nu(\alpha) = \Phi^\nu(\beta, \gamma).$$

Die Berechnung neuer Näherungswerte für die Parameter wird in den zwei Gruppen unterschiedlich vollzogen. Bei der „modifizierten Iteration" ersetzen wir den Teilschritt 3 aus der „einfachen Iteration" durch die Vorschrift:

3′. Es werden α^ν und $\Delta\alpha$ gemäß $\alpha^\nu = (\beta^\nu, \gamma^\nu)$ bzw. $\Delta\alpha = (\Delta\beta, \Delta\gamma)$ aufgespalten, und es wird $t = 1$ gesetzt. Für β werden die Werte $\tilde{\tilde{\beta}}$ ermittelt, bei denen

$$\Phi^\nu(\beta, \gamma^\nu + t\,\Delta\gamma)$$

minimiert wird, und es wird getestet, ob

$$\Phi(\tilde{\tilde{\beta}}, \gamma^\nu + t\,\Delta\gamma) \leqq \Phi(\alpha^\nu) - \tfrac{1}{3} t\,\Delta\Phi \tag{10′}$$

gilt. Sofern die Bedingung (10′) erfüllt ist, wird ein neuer Iterationsschritt mit $\alpha^{\nu+1} = (\tilde{\tilde{\beta}}, \gamma^\nu + t\,\Delta\gamma)$ begonnen. Andernfalls wird der Faktor t halbiert und der Teilschritt 3′ wiederholt.

Bei der erweiterten Vorschrift wird nicht nur mit jedem *Iterationsschritt* ein Minimumproblem gelöst, sondern auch mit jedem *Dämpfungsschritt*. Allerdings ist die Anzahl der freien Parameter bei den Dämpfungsschritten reduziert, und der Arbeitsaufwand ist dementsprechend kleiner. Andererseits gilt offensichtlich

$$\Phi^\nu(\tilde{\tilde{\beta}}, \gamma^\nu + t\,\Delta\gamma) \leqq \Phi^\nu(\beta^\nu + t\,\Delta\beta, \gamma^\nu + t\,\Delta\gamma),$$

und so ist für manche Werte von t wohl die Bedingung (10′) aber nicht (10) erfüllt. Für die Iteration nach der ursprünglich gegebenen Vorschrift wären deshalb noch weitere Dämpfungsschritte erforderlich, ehe eine bessere Näherungslösung gefunden ist. Damit wird der Mehraufwand bei der Ausführung von 3′ gerechtfertigt.

4

Es scheint zunächst so, als ob durch die Aufteilung der Parameter in zwei Gruppen und durch die verschiedene Behandlung eine Symmetrie der klassischen Verfahren zerstört wird. Aber es zeigt sich im Gegenteil, daß bei der Anpassung mit Exponentialsummen

$$F(x, \alpha) = \sum_{k=1}^{l} \beta_k e^{\gamma_k x} \tag{14}$$

gerade durch die Modifikation eine Symmetrie zwischen den Parametern β_k und γ_k erreicht wird. In Hinblick darauf diskutieren wir das Verhalten der Iterationsfolgen bei Verschiebung des Approximationsintervalls.

Über einem reellen Intervall X sei die stetige Funktion $f(x)$ durch Exponentialsummen (14) im Tschebyscheffschen Sinn bestmöglich zu approximieren. Sei ξ eine feste reelle Zahl und

$$\hat{X} = \{\hat{x} \mid \hat{x} = x + \xi,\ x \in X\}$$

das transformierte Intervall. Wir führen einen Verschiebungsoperator

$$P\colon C(X) \to C(\hat{X})$$

durch

$$(Pf)(x) = f(x - \xi) \tag{15}$$

ein. P bildet die Familie der Exponentialsummen in sich ab. Offensichtlich gilt: Wenn $F[\alpha]$ beste Approximation zu f im Intervall X ist, dann ist auch $PF[\alpha]$ beste Approximation zu Pf im Intervall $\hat{X}$.

Definition. Sei $F[\alpha^\nu]$ eine Iterationsfolge zur Konstruktion der besten Approximation zu $f \in C(X)$ bei der Anpassung in X, und $F[\hat{\alpha}^\nu]$ sei die Folge zu Pf bei der Anpassung in $\hat{X}$. Die Iteration heißt invariant gegenüber Translationen, wenn aus

$$F[\hat{\alpha}^0] = PF[\alpha^0]$$

die Relationen

$$F[\hat{\alpha}^\nu] = PF[\alpha^\nu] \qquad \text{für } \nu = 1, 2, 3 \ldots$$

folgen.

Sei nun $F(x, \alpha^0) = \sum_{k=1}^{l} \beta_k e^{\gamma_k x}$ die Ausgangsnäherung und $\varepsilon(x) = f(x) - F(x, \alpha^0)$ die Fehlerfunktion. Nach (2) und (6) bedeutet die Minimierung von $\Phi^0(\alpha)$ die Bestimmung von Änderungen $\Delta\alpha_k$, so daß an den Punkten x_i für die Differenz

$$\varepsilon(x) - \sum_{k=1}^{n} \Delta\alpha_n^0 \frac{\partial}{\partial \alpha_k} F(x, \alpha^0)$$

das Maximum der Beträge möglichst klein wird. Wie man durch Ausrechnen der Ableitungen für die Exponentialsummen (14) erkennt, ist dazu die Funktion $e^0(x)$

der Form

$$e^0(x) = \sum_{k=1}^{l} (\sigma_k + \tau_k x)\, e^{\gamma_k x}$$

mit den freuen Parametern σ_k, τ_k aufzusuchen, welche die Fehlerkurve $\varepsilon(x)$ am besten approximiert. Aus $e^0(x)$ ergeben sich die Parameteränderungen

$$\Delta\beta_k = \sigma_k, \qquad \Delta\gamma_k = \tau_k/\beta_k \qquad (\beta_k \neq 0),\ k = 1, 2, \ldots, l. \tag{17}$$

Für das transformierte Problem sei (16) erfüllt. Dann hat

$$F[\hat{\alpha}^0] = PF[\alpha^0] = \sum_{k=1}^{l} \hat{\beta}_k\, e^{\hat{\gamma}_k x} = \sum_{k=1}^{l} (\beta_k e^{-\gamma_k \xi})\, e^{\gamma_k x} \tag{18}$$

die gleichen Frequenzen γ_k wie $E[\alpha^0]$ und die zugehörige Fehlerkurve ist $\varepsilon[\hat{\alpha}^0] = P\varepsilon[\alpha^0]$. Wir können annehmen, daß beim linearisierten Problem die Anpassung auf den um die Größe ξ verschobenen Punkten

$$\hat{x}_i = x_i + \xi \in \hat{X}$$

geschieht; denn die Wahl hängt bei Remes-Algorithmen [7] nur von der Gestalt der Fehlerkurve $\varepsilon(x)$ ab. Also lautet die Lösung des zweiten Teilschrittes

$$\hat{e}^0(x) = (P e^0)(x) = \sum_{k=1}^{l} (\hat{\sigma}_k + \hat{\tau}_k x)\, e^{\gamma_k x}$$

mit

$$\hat{\sigma}_k = (\sigma_k - \tau_k \xi)\, e^{-\gamma_n \xi}, \qquad \hat{\tau}_k = \tau_k e^{-\gamma_n \xi}. \tag{19}$$

Aus (17), (18) und (19) folgt unmittelbar, daß bei beiden Iterationen die Frequenzänderungen übereinstimmen, daß sich aber die Änderungen der Faktoren β_k nicht entsprechen. Die einfache Iteration führt also nicht auf invariante Folgen. Bei der modifizierten Iteration werden dagegen zu den vorgegebenen Frequenzen jeweils die optimalen Faktoren β_k eingesetzt. Da dieser Vorgang invariant gegenüber den Verschiebungen ist, gilt das auch für die ganze Iterationsfolge.

5

Es seien noch zwei Beispiele für Iterationsverfahren genannt, bei denen das beschriebene Konstruktionsprinzip angewandt wird. Es genügt, für die Beispiele das Funktional Φ zu definieren und für den 2. Teilschritt die spezielle Vorschrift zur Berechnung von $\Delta\alpha$ und $\Delta\Phi$ anzugeben. Die anderen Teilschritte sind schon durch den allgemeinen Formalismus vollständig festgelegt.

Beispiel 1. Anpassung nach der Methode der kleinsten Quadrate [1, 4, 7]. Es sind für den Funktionsansatz $F(x, \alpha)$ die Parameter α_k so zu bestimmen, daß die gewichtete Fehlerquadratsumme

$$\Phi(\alpha) = \sum_{i=1}^{m} w_i [F(x_i, \alpha) - y_i]^2, \qquad w_i > 0,\ m \geqq n \tag{20}$$

minimal wird. Es wird $\Phi^\nu(\alpha) = \Phi(\alpha)$ gesetzt und $\Delta\alpha$ ergibt sich aus dem Gleichungssystem

$$J' W J \Delta\alpha - J' W h = 0.$$

Dabei berechnet sich die Jakobische Matrix J, die Wichtematrix W und der Fehlervektor h gemäß

$$J_{ik} = \frac{\partial}{\partial \alpha_k} F(x_i, \alpha^\nu),$$
$$W_{ii} = w_i, \qquad W_{ik} = 0 \qquad \text{für } i \neq k,$$
$$h_i = y_i - F(x_i, \alpha^\nu).$$

Ferner gilt $\Delta\Phi = \Delta\alpha' J' W h$.

Beispiel 2. Lösung nichtlinearer Gleichungssysteme [3, 5]. Eine Bewertung von Näherungslösungen kann man wie in [5] durch die Fehlerquadratsumme

$$\Phi(\alpha) = \sum_{i=1}^{n} w_i [F_i(\alpha)]^2$$

oder durch das Fehlermaximum

$$\Phi(\alpha) = \max_{1 \leqq i \leqq n} |F_i(\alpha)|$$

einführen. Dann ist auch $\Phi^\nu(\alpha) = \Phi(\alpha)$ möglich. In beiden Fällen führt die Minimierung im zweiten Teilschritt auf das zu (3) äquivalente Gleichungssystem $\nabla F[\alpha^\nu] \cdot \Delta\alpha = -F[\alpha^\nu]$. Wegen $\tilde{\Phi}^\nu(\alpha^\nu + \Delta\alpha) = 0$ vereinfacht sich die Forderung (10) zu

$$\Phi(\alpha^\nu + t\,\Delta\alpha) \leqq \left(1 - \frac{t}{3}\right) \Phi(\alpha^\nu),$$

und eine entsprechende Vereinfachung ergibt sich für (10').

6

Zum Schluß betrachten wir noch ein numerisches Beispiel. Die Untersuchung der Reaktionsgeschwindigkeit bei bestimmten chemischen Prozessen führt auf die Differentialgleichung

$$\frac{dy}{du} = \frac{y}{a u - b y + c}, \qquad u > 0.$$

Die Parameter a, b und c sind aus den Meßwerten zu bestimmen. Die Lösungen der Differentialgleichung haben die Gestalt

$$y = \frac{c}{b} + \frac{a}{1+b} e^{-x} + k \cdot e^{bx} \quad \text{mit} \quad u = e^{-x}.$$

Dies legt den Ansatz

$$y = \alpha_1 + \alpha_2 e^{-x} + \alpha_3 e^{\alpha_4 x}$$

nahe. Jedoch wird dann die Jakobische Matrix für $\alpha_4 \to 0$ singulär. Deshalb ist der äquivalente Ansatz

$$y = \alpha_1 + \alpha_2 e^{-x} + \alpha_3 \frac{e^{\alpha_4 x} - 1}{\alpha_4}$$

geeigneter. Der Einfluß der Beschleunigung zeigt sich nun deutlich bei der Minimierung der Fehlerquadratsumme (20) für die Punkte (Testwerte)

$$x_i = i, \qquad y_i = \sqrt{5i}, \qquad i = 1, 2 \ldots 7$$

mit den Gewichten $W_i = 1$. Als optimale Anpassung erhalten wir die Funktion

$$y = 1{,}91 - 1{,}10e^{-x} + 0{,}763\,\frac{e^{-0{,}0867x} - 1}{-0{,}0867}\,.$$

Die herkömmliche Iteration verlangt 30 Schritte. Wenn man aber ausnutzt, daß die Funktion in dem ersten (bzw. in den ersten beiden, bzw. in den ersten drei) Argumenten linear ist, reduziert sich die Schrittzahl auf 21 (bzw. 18, bzw. 12).

Herrn Professor Dr. H. Werner möchte ich für seine Anregungen danken.

Literatur

1. Braess, D.: Über Dämpfung bei Minimalisierungsverfahren. Computing **1**, 264—272 (1966).
2. — Die Konstruktion der Tschebyscheff-Approximierenden bei der Anpassung an Exponentialsummen, J. Approx. Theory 1970. S.a.: Approximation mit Exponentialsummen. Habilitationsschrift Münster 1967.
3. —, u. Späth, H.: Maßnahmen zur globalen Konvergenzerzwingung beim Newtonschen Verfahren für spezielle nichtlineare Gleichungssysteme. ZAMM **47**, 409—410 (1967).
4. Fletcher, R., Powell, J.M.D.: A rapidly convergent method for minimization. Comp. J. **6**, 163—168 (1963).
5. Gleyzal, A.N.: Solution of non-linear equations. Quart. Appl. Math. **17**, 95 (1959).
6. Marquardt, D.W.: An algorithm for least-squares estimation of non-linear parameters. J. Soc. Indust. Appl. Math. **11**, 431—441 (1963).
7. Meinardus, G.: Approximation of functions: Theory and numerical methods. S. 105 und 149. Berlin-Heidelberg-New York: Springer 1967.
8. Ostrowski, A.M.: Contributions to the theory of the method of steepest descent. Arch. Rational Mech. Anal. **26**, 257—280 (1967).
9. Urabe, M.: Convergence of numerical iteration in solution of equations. Journal Science Hiroshima University A **19**, 479—489 (1956).
10. Werner, H.: Die konstruktive Ermittlung der Tschebyscheff-Approximierenden im Bereich der rationalen Funktionen. Arch. Rational Mech. Anal. **11**, 368—384 (1962).
11. Wetterling, W.: Anwendung des Newtonschen Iterationsverfahrens bei der Tschebyscheff-Approximation, insbesondere mit nichtlinear auftretenden Parametern. MTW **10**, 61—63, 112—115 (1963).

Dr. Dietrich Braess
Institut f. Numerische u.
instrumentelle Mathematik
d. Universität
44 Münster, Roxeler Str.

Numer. Math. 14, 476—494 (1970)

Existenz und Eindeutigkeit für Lösungen nichtlinearer Randwertprobleme

Walter Petry

Eingegangen am 19. Juni 1968

1. Einleitung

Ehrmann [3] und Conti [1] geben Existenzsätze für die Lösung des Problems

$$L_1 u = M_1 u, \tag{1.1a}$$

$$L_2 u = M_2 u \tag{1.1b}$$

an, wobei L_1 und L_2 lineare Abbildungen eines Banachraumes B_0 in Banachräume B_1 bzw. B_2 und M_1 und M_2 (nichtlineare) Abbildungen von B_0 in B_1 bzw. B_2 sind. Hierbei wird vorausgesetzt, daß die Operatoren M_1 und M_2 auf dem ganzen Raum B_0 beschränkt sind, d.h. es gibt positive Konstanten a_1, a_2, so daß für alle $u \in B_0$ gilt

$$\|M_1 u\|_{B_1} \leqq a_1, \quad \|M_2 u\|_{B_2} \leqq a_2. \tag{1.2}$$ [1]

In einer späteren Arbeit von Ehrmann [4] wird die Beschränktheitsbedingung (1.2) durch eine Bedingung der Form

$$\|M_1 u\|_{B_1} \leqq a_1 + b_1 \|u\|_{B_0}, \quad \|M_2 u\|_{B_2} \leqq a_2 + b_2 \|u\|_{B_0} \tag{1.3}$$

ersetzt, wobei b_i $(i = 1, 2)$ nichtnegative Konstanten sind, die gewissen Bedingungen genügen müssen.

In dieser Arbeit werden unter der Voraussetzung, daß der durch $L_1 u = \Theta_{B_1}$, $L_2 u = \Theta_{B_2}$ [2] erzeugte lineare Operator formal selbstadjungiert ist, zwei Existenzsätze und zwei Existenz- und Eindeutigkeitssätze für „verallgemeinerte" Lösungen des Problems (1.1) bewiesen, welche die lineare Beschränktheit von M_1 nicht benötigen.

Diese Sätze ergeben sich aus einem abstrakten Existenzsatz und einem abstrakten Existenz- und Eindeutigkeitssatz des Verfassers [8].

Als Anwendung der erhaltenen Ergebnisse betrachten wir für eine offene beschränkte Menge $\Omega \subset R^2$ die Differentialgleichung

$$-\Delta u(s) = f\big(s, u(s)\big) \quad \text{für} \quad s \in \Omega \tag{1.4}$$

mit gewissen nichtlinearen Randbedingungen für $u(s)$ und mit auf $\mathfrak{H} = \{(s, u) \mid s \in \bar{\Omega}, u \in R^1\}$ stetigem $f(s, u)$. Ebenso untersuchen wir die Differentialgleichung

$$-u''(s) = g\big(s, u(s)\big) \quad \text{für} \quad s \in]0, 1[\tag{1.5}$$

1 $\|\cdot\|_B$ bezeichnet die Norm von B.

2 Θ_B bezeichnet das Nullelement von B.

mit gewissen nichtlinearen Randbedingungen, wobei $g(s, u)$ am Rand Singularitäten besitzen darf. Für $s \in]0, 1[$ sei $g(s, u)$ stetig.

Die Sätze von Ehrmann und Conti sind auf diese Beispiele nicht anwendbar.

2. Hilfsmittel

In diesem Abschnitt formulieren wir zwei abstrakte Sätze, die in einer früheren Arbeit des Verfassers [8] bewiesen wurden, und welche beim Beweis der Existenz- und Eindeutigkeitssätze für das Problem (1.1) benötigt werden.

Bezeichnung 1. Im folgenden bezeichne H einen Hilbertraum mit dem Skalarprodukt $(\cdot, \cdot)$ und B einen Banachraum. Ferner sei für $0 \leqq \alpha < \infty$ $K_\alpha = \{y \mid y \in B, \|y\|_B \leqq \alpha\}$ und $L_\alpha = \{x \mid x \in H, \|x\|_H \leqq \alpha\}$.

Hilfssatz 1. Es seien die folgenden Voraussetzungen erfüllt.

(A) Es existiere ein R mit $0 \leqq R < \infty$, so daß gelte:

(α) $T(\cdot, \cdot)$ sei eine stetige Abbildung von $H \times K_R$ in B.

(β) Für $y \in K_R$ sei $S(\cdot, y)$ eine stetige Abbildung von H in sich.

(γ) Für $x \in H$ sei $S(x, \cdot)$ eine stetige Abbildung von K_R in H.

(B) Für $y \in K_R$ und $x_1, x_2 \in H$ gelte

$$\operatorname{Re}\left(S(x_1, y) - S(x_2, y), x_1 - x_2\right) \geqq C \|x_1 - x_2\|_H^2 \tag{2.1}$$

mit einer Konstanten $C > 0$.

(C) Für $y \in K_R$ gelte

$$\|S(\Theta_H, y)\|_H \leqq M \tag{2.2}$$

mit einer Konstanten $M > 0$.

(D) Für $x \in L_{M/C}$ und $y \in K_R$ gelte

$$\|T(x, y)\|_B \leqq R. \tag{2.3}$$

(E) $T(\cdot, \cdot)$ sei eine kompakte Abbildung.

Dann existiert mindestens ein x^*, y^* mit $x^* \in L_{M/C}$ und $y^* \in K_R$ von

$$S(x, y) = \Theta_H, \quad y = T(x, y). \tag{2.4}$$

Bemerkung 1. Setzt man $S(x, y) = x - S_0(x, y)$, dann kann das Problem (2.4) in der Form

$$x = S_0(x, y), \quad y = T(x, y) \tag{2.5}$$

geschrieben werden.

Hilfssatz 2. Es seien die folgenden Voraussetzungen erfüllt.

(A') Es existieren Konstanten R und M mit $0 \leqq R < \infty$ und $0 \leqq M < \infty$, so daß gelte:

(α) $T(\cdot, \cdot)$ sei eine stetige Abbildung von $L_{M/C} \times K_R$ in B.

(β) Wie (Iβ) von Hilfssatz 1.

(γ) Es existiere eine Konstante $L > 0$, so daß für $x \in L_{M/C}$ und $y_1, y_2 \in K_R$ gelte

$$\|S_0(x, y_1) - S_0(x, y_2)\|_H \leqq L \|y_1 - y_2\|_B. \tag{2.6}$$

(B′) Wie (B) von Hilfssatz 1.

(C′) Wie (C) von Hilfssatz 1.

(D′) Wie (D) von Hilfssatz 1.

(E′) Für $x_i \in L_{M/C}$ und $y_i \in K_R$ $(i=1, 2)$ gelte

$$\|T(x_1, y_1) - T(x_2, y_2)\|_B \leqq k_1 \|x_1 - x_2\|_H + k_2 \|y_1 - y_2\|_B \tag{2.7}$$

mit Konstanten $k_i \geqq 0$ $(i=1, 2)$ und

$$r := k_1 \frac{L}{C} + k_2 < 1. \tag{2.8}$$

Dann besitzt das System

$$x = S_0(x, y), \quad y = T(x, y) \tag{2.9}$$

genau eine Lösung x^*, y^* mit $y^* \in K_R$. Es ist $x^* \in L_{M/C}$. Ferner konvergiert für $y_0 \in K_R$ das Iterationsverfahren

$$x_{\nu+1} = S_0(x_{\nu+1}, y_\nu), \quad y_{\nu+1} = T(x_{\nu+1}, y_\nu) \quad (\nu = 0, 1, 2 \ldots) \tag{2.10}$$

gegen die eindeutige Lösung x^*, y^*.

3. Existenz- und Eindeutigkeitssätze

In diesem Abschnitt untersuchen wir Existenz und Eindeutigkeit einer „verallgemeinerten" Lösung der nichtlinearen Randwertaufgabe

$$(L_1 u)(s) = (M_1 u)(s) \quad \text{für} \quad s \in \Omega, \tag{3.1a}$$

$$L_2 u = M_2 u. \tag{3.1b}$$

Hierbei bezeichnet Ω eine offene, beschränkte, meßbare Menge aus R^l $(l \geqq 1)$ mit dem Maß $|\Omega|$.

Für die anschließenden Betrachtungen werden folgende Räume eingeführt:

E^n: Raum der reellen n-dimensionalen Vektoren mit dem Skalarprodukt $\langle z_1, z_2 \rangle = \sum_{i=1}^{n} \overset{i}{z}_1 \cdot \overset{i}{z}_2$ und der Norm $\|z\|_{E^n} = (\langle z, z \rangle)^{\frac{1}{2}}$ (Hilbertraum).

$L^p[\Omega]$: Menge der auf Ω reellen meßbaren Funktionen $z(s)$ mit $\int_\Omega |z(s)|^p \, ds < \infty$ $(1 \leqq p < \infty)$ und der Norm $\|z\|_{L^p} = \left(\int_\Omega |z(s)|^p \, ds \right)^{1/p}$. Hierbei werden Funktionen, die fast überall auf Ω gleich sind, als identisch betrachtet.

$\mathscr{L}^{\mathfrak{p}}$: Es sei $\mathfrak{p} = (p_1, \ldots, p_n)$, dann sei $\mathscr{L}^{\mathfrak{p}} = \prod_{i=1}^{n} L^{p_i}[\Omega]$ mit der Norm $\|u\|_{\mathscr{L}^{\mathfrak{p}}} = \left(\sum_{i=1}^{n} \|\overset{i}{u}\|^2_{L^{p_i}} \right)^{\frac{1}{2}}$ (Banachraum).

Für $p_i = 2$ $(i = 1, \ldots, n)$ erhält man speziell einen Hilbertraum, der im folgenden mit $\mathscr{L}^2$ bezeichnet wird.

Voraussetzung 1

(I) Es sei L_1 ein linearer Operator mit Definitionsbereich $D(L_1) \subset \mathscr{L}^{\mathfrak{p}}$ mit $p_i \geqq 2$ $(i = 1, \ldots, n)$ und Wertebereich $R(L_1)$ aus einem linearen Funktionenraum $\tilde{B}$. M_1 sei ein (nichtlinearer) stetiger beschränkter Operator von $\mathscr{L}^{\mathfrak{p}}$ in $\mathscr{L}^{\mathfrak{q}}$ mit $\mathfrak{q} =$

$(q_1, \ldots, q_n)$ und $\frac{1}{p_i} + \frac{1}{q_i} = 1$ $(i = 1, \ldots, n)$. L_2 sei ein linearer Operator mit $D(L_1) \subset D(L_2) \subset \mathscr{L}^{\mathfrak{p}}$ und $R(L_2)$ aus einem Banachraum B. M_2 sei ein (nichtlinearer) stetiger beschränkter Operator von $\mathscr{L}^{\mathfrak{p}}$ in B.

(II) Sei $\tilde{B}_0$ eine nichtleere Teilmenge von $\mathscr{L}^{\mathfrak{q}} \cap \tilde{B}$ und sei $f \in \tilde{B}_0$, dann existiere eine Abbildung K, so daß aus

$$u(s) = (Kf)(s) \qquad (s \in \Omega) \tag{3.2a}$$

folgt $u \in D(L_1)$ und

$$(L_1 u)(s) = f(s) \quad (s \in \Omega), \qquad L_2 u = \Theta_B. \tag{3.2b}$$

Die Abbildung K sei linear und beschränkt von $\mathscr{L}^{\mathfrak{q}}$ in $\mathscr{L}^{\mathfrak{p}}$. Aufgefaßt als Abbildung auf $\mathscr{L}^2$ sei K selbstadjungiert und positiv definit.

(III) Für $z \in \mathscr{L}^{\mathfrak{q}}$ gelte die Darstellung

$$Kz = \mathscr{H}\mathscr{H}^* z. \tag{3.3}$$

Hierbei sei $\mathscr{H}$ ein linearer beschränkter Operator von $\mathscr{L}^2$ in $\mathscr{L}^{\mathfrak{p}}$ und $\mathscr{H}^*$ der zu $\mathscr{H}$ adjungierte Operator, d.h. $\mathscr{H}^*$ ist linear und beschränkt von $\mathscr{L}^{\mathfrak{q}}$ in $\mathscr{L}^2$.

(IV) Für $y \in B$ existiere $u_0(y) \in D(L_1)$ mit

$$\big(L_1 u_0(y)\big)(s) \equiv 0 \quad \text{für} \quad s \in \Omega. \tag{3.4}$$

Hierbei sei u_0 ein linearer beschränkter Operator von B in $\mathscr{L}^{\mathfrak{p}}$.

(V) Es existiere auf B ein linearer beschränkter Operator $\tilde{L}$ mit

$$L_2 u_0(\tilde{L} y) = \tilde{L}(L_2 u_0)\, y = y, \tag{3.5}$$

d.h. $\tilde{L} = (L_2 u_0)^{-1}$.

(VI) Für $u_j \in \mathscr{L}^{\mathfrak{p}}$ $(j = 1, 2)$ gelte

$$\int_\Omega \langle M_1(u_1)(s) - M_1(u_2)(s),\, u_1(s) - u_2(s) \rangle \, ds \leqq a \, \|u_1 - u_2\|^2_{\mathscr{L}^2}. \tag{3.6}$$ [3]

Hierbei sei

$$a \geqq 0, \qquad C := 1 - a \, \|\mathscr{H}\|^2_{\mathscr{L}^2 \to \mathscr{L}^{\mathfrak{p}}} > 0. \tag{3.7}$$ [4]

(VII) Es existiere eine Konstante $b > 0$, so daß für $u \in \mathscr{L}^{\mathfrak{p}}$ gelte

$$\|M_2 u\|_B \leqq b. \tag{3.8}$$

(VIII) $\tilde{L} M_2$ sei kompakt von $\mathscr{L}^{\mathfrak{p}}$ in B.

Voraussetzung 2

(I') Es sei L_1 ein linearer Operator mit $D(L_1) \subset B_1 \subset \mathscr{L}^2$ und $R(L_1)$ aus einem linearen Funktionenraum $\tilde{B}$. M_1 sei ein (nichtlinearer) stetiger und beschränkter Operator von B_1 in $B_2 \subset \mathscr{L}^2$. L_2 sei ein linearer Operator mit $D(L_1) \subset D(L_2) \subset B_1$ und $R(L_2)$ aus einem Banachraum B. M_2 sei ein (nichtlinearer) stetiger beschränkter Operator von B_1 in B.

3 Es ist $\mathscr{L}^{\mathfrak{p}} \subset \mathscr{L}^2$ wegen $p_i \geqq 2$ und $|\Omega| < \infty$.

4 $\|\alpha\|_{\beta \to \gamma}$ bezeichnet die Operatornorm von α als Abbildung von β in γ.

(II′) Sei $\tilde{B}_0$ eine nichtleere Teilmenge von $B_2 \cap \tilde{B}$ und sei $f \in \tilde{B}_0$, dann existiere eine Abbildung K, so daß aus

$$u(s) = (Kf)(s) \quad (s \in \Omega) \tag{3.9a}$$

folgt $u \in D(L_1)$ und

$$(L_1 u)(s) = f(s) \quad (s \in \Omega), \qquad L_2 u = \Theta_B. \tag{3.9b}$$

Hierbei sei K ein linearer beschränkter Operator von B_2 in B_1. Es existiere ein linearer, beschränkter, selbstadjungierter, positiv definiter Operator $\tilde{K}$ auf $\mathscr{L}^2$ mit $J_1 K = \tilde{K} J_2$, wobei J_i $(i = 1, 2)$ die Injektionsabbildungen von B_i in $\mathscr{L}^2$ bedeuten.

(III′) Die Operatoren $K_1^{\frac{1}{2}}$ und $K_2^{\frac{1}{2}}$ seien linear und beschränkt von $\mathscr{L}^2$ bzw. B_2 in B_1 bzw. $\mathscr{L}^2$ und es gelte $K^{\frac{1}{2}} = J_1 K_1^{\frac{1}{2}}$ bzw. $K_2^{\frac{1}{2}} = K^{\frac{1}{2}} J_2$, wobei $K^{\frac{1}{2}}$ den eindeutig bestimmten selbstadjungierten, positiv definiten Wurzeloperator von $\tilde{K}$ bezeichnet.

(IV′) Für $y \in B$ existiere $u_0(y) \in D(L_1)$ mit

$$\big(L_1 u_0(y)\big)(s) \equiv 0 \quad \text{für} \quad s \in \Omega. \tag{3.10}$$

Hierbei sei u_0 ein linearer beschränkter Operator von B in B_1.

(V′) Wie (V).

(VI′) Für $u_j \in B_1$ $(j = 1, 2)$ gelte

$$\big(J_2 M_1(u_1) - J_2 M_1(u_2),\, J_1 u_1 - J_1 u_2\big) \leqq a \,\|J_1 u_1 - J_1 u_2\|^2_{\mathscr{L}^2}. \tag{3.11}$$

Hierbei bezeichne $(\cdot, \cdot)$ das Skalarprodukt in $\mathscr{L}^2$, und es sei

$$a \geqq 0, \qquad C := 1 - a\,\|K^{\frac{1}{2}}\|^2_{\mathscr{L}^2 \to \mathscr{L}^2} > 0. \tag{3.12}$$

(VII′) Es existiere eine Konstante $b > 0$, so daß für $u \in B_1$ gelte

$$\|M_2 u\|_B \leqq b. \tag{3.13}$$

(VIII′) $\tilde{L} M_2$ sei kompakt von B_1 in B.

Lemma 1. Es sei die Voraussetzung 1 erfüllt. Ferner sei $v^* \in \mathscr{L}^2$, $y^* \in B$ Lösung von

$$\begin{aligned} v &= \mathscr{H}^* M_1\big(u_0(y) + \mathscr{H} v\big), \\ y &= \tilde{L} M_2\big(u_0(y) + \mathscr{H} v\big). \end{aligned} \tag{3.14}$$

Dann gilt

$$u^* := u_0(y^*) + \mathscr{H} v^* \in \mathscr{L}^{\mathfrak{p}} \tag{3.15}$$

und u^*, y^* genügt den Operatorgleichungen

$$\begin{aligned} u &= u_0(y) + K M_1(u) \\ L_2 u_0(y) &= M_2 u. \end{aligned} \tag{3.16}$$

Beweis. Die Behauptung folgt aus (3.14) unter Beachtung von (IV), (I), (III) und (V).

Lemma 2. Es sei die Voraussetzung 2 erfüllt. Ferner sei $v^* \in \mathscr{L}^2$, $y^* \in B$ Lösung von

$$\begin{aligned} v &= K_2^{\frac{1}{2}} M_1 \left(u_0(y) + K_1^{\frac{1}{2}} v\right), \\ y &= \tilde{L} M_2 \left(u_0(y) + K_1^{\frac{1}{2}} v\right). \end{aligned} \tag{3.17}$$

Dann gilt

$$u^* := u_0(y^*) + K_1^{\frac{1}{2}} v^* \in B_1 \tag{3.18}$$

und u^*, y^* genügt den Operatorgleichungen (3.16).

Beweis. Die Behauptung folgt aus (3.17) unter Beachtung von (IV'), (I'), (V') und der aus (II') und (III') folgenden Bezeichnung $K = K_1^{\frac{1}{2}} K_2^{\frac{1}{2}}$.

Bemerkung 2. Unter der Voraussetzung 1 (Voraussetzung 2) gelte für die Lösung u^*, y^* von (3.16)

$$M_1(u^*) \in \tilde{B}_0, \qquad K M_1(u^*) \in D(L_1). \tag{3.19}$$

Dann ist u^* eine „klassische" Lösung von (3.1).

Beweis. Nach Voraussetzung gilt

$$\begin{aligned} u^* &= u_0(y^*) + K M_1(u^*) \\ L_2 u_0(y^*) &= M_2 u^*. \end{aligned} \tag{3.20}$$

Nun ist nach (IV) bzw. (IV') $u_0(y^*) \in D(L_1)$ und es gilt

$$\left(L_1 u_0(y^*)\right)(s) \equiv 0 \quad \text{für} \quad s \in \Omega.$$

Aus (3.20) folgt daher mittels (3.19) unter Beachtung von (II) bzw. (II') $u^* \in D(L_1)$ mit

$$\begin{aligned} (L_1 u^*)(s) &= \left(L_1 K M_1(u^*)\right)(s) = M_1(u^*)(s) \qquad (s \in \Omega), \\ L_2 u^* &= L_2 u_0(y^*) + L_2 K M_1(u^*) = L_2 u_0(y^*) = M_2 u^*. \end{aligned}$$

Damit ist Bemerkung 2 bewiesen.

Definition 1. Sei u^*, y^* mit $u^* \in \mathscr{L}^p$ bzw. $u^* \in B_1$, $y^* \in B$ Lösung von (3.16), dann heißt u^* eine „verallgemeinerte" Lösung von (3.1).

Aus dieser Definition und Bemerkung 2 folgt

Lemma 3. Unter der Voraussetzung 1 bzw. Voraussetzung 2 sei u^* eine „verallgemeinerte" Lösung von (3.1) und es gelte (3.19). Dann ist u^* eine „klassische" Lösung von (3.1).

Im folgenden werden daher nur Existenzsätze für „verallgemeinerte" Lösungen bewiesen. Es gilt der

Satz 1. Die Voraussetzung 1 sei erfüllt. Dann gibt es mindestens eine „verallgemeinerte" Lösung von (3.1).

Beweis. Nach Lemma 1 genügt es die Existenz einer Lösung $v^* \in \mathscr{L}^2$, $y^* \in B$ von (3.14) nachzuweisen. Wir wenden Hilfssatz 1 an und setzen $B = B$, $H = \mathscr{L}^2$, $R = b \|\tilde{L}\|_{B \to B}$, $S(v, y) = v - \mathscr{H}^* M_1\left(u_0(y) + \mathscr{H} v\right)$, $T(v, y) = \tilde{L} M_2\left(u_0(y) + \mathscr{H} v\right)$.

Überprüfung der Voraussetzungen des Hilfssatzes:

(Aα) $T(\cdot,\cdot)$ ist nach (III), (IV), (I) und (V) stetig von $\mathscr{L}^2 \times B$ in B.

(Aβ,γ) $S(\cdot,\cdot)$ ist nach (III), (IV) und (I) stetig von $\mathscr{L}^2 \times B$ in $\mathscr{L}^2$, woraus (Aβ) und (Aγ) folgen.

(B) Sei $y \in B$ und seien $v_1, v_2 \in \mathscr{L}^2$, dann folgt aus (VI)

$$\begin{aligned}(S(v_1, y) - S(v_2, y), v_1 - v_2) &= \|v_1 - v_2\|^2_{\mathscr{L}^2} \\ &\quad - (\mathscr{H}^* (M_1(u_0(y) + \mathscr{H} v_1) - M_1(u_0(y) + \mathscr{H} v_2)), v_1 - v_2) \geqq \|v_1 - v_2\|^2_{\mathscr{L}^2} \\ &\quad - \int_\Omega \langle (M_1(u_0(y) + \mathscr{H} v_1) - M_1(u_0(y) + \mathscr{H} v_2))(s), (\mathscr{H}(v_1 - v_2))(s) \rangle \, ds \\ &\geqq \|v_1 - v_2\|^2_{\mathscr{L}^2} - a \|\mathscr{H}(v_1 - v_2)\|^2_{\mathscr{L}^2} \geqq (1 - a \|\mathscr{H}\|^2_{\mathscr{L}^2 \to \mathscr{L}^2}) \|v_1 - v_2\|^2_{\mathscr{L}^2} \\ &\geqq C \|v_1 - v_2\|^2_{\mathscr{L}^2},\end{aligned}$$

womit Voraussetzung (B) bewiesen ist.

(C) Sei $y \in B$ mit $\|y\|_B \leqq R$, dann folgt aus

$$S(\Theta_{\mathscr{L}^2}, y) = \mathscr{H}^* M_1(u_0(y))$$

nach (IV), (I) und (III) Voraussetzung (C).

(D) Sei $v \in \mathscr{L}^2$ und $y \in B$, dann folgt nach (VII)

$$\|T(v, y)\|_B \leqq \|\tilde{L}\|_{B \to B} \|M_2(u_0(y) + \mathscr{H} v)\|_B \leqq b \|\tilde{L}\|_{B \to B} = R.$$

(E) Sei $v \in \mathscr{L}^2$, $y \in B$ mit $\|v\|_{\mathscr{L}^2} \leqq A_1$, $\|y\|_B \leqq A_2$, dann existiert eine Konstante $A_3 > 0$, so daß $\|u_0(y) + \mathscr{H} v\|_{\mathscr{L}^p} \leqq A_3$ gilt. Nach Voraussetzung (VIII) ist dann $T(\cdot,\cdot)$ kompakt von $\mathscr{L}^2 \times B$ in B.

Die Voraussetzungen von Hilfssatz 1 sind daher erfüllt. Es existiert daher mindestens ein $v^* \in \mathscr{L}^2$, $y^* \in B$ von (3.14), womit Satz 1 bewiesen ist.

Ein analoger Satz gilt unter der Voraussetzung 2.

Satz 2. Es sei Voraussetzung 2 erfüllt. Dann gibt es mindestens eine „verallgemeinerte" Lösung von (3.1).

Beweis. Nach Lemma 2 genügt es die Existenz einer Lösung $v^* \in \mathscr{L}^2$, $y^* \in B$ von (3.17) nachzuweisen. Dazu wendet man wieder Hilfssatz 1 an und setzt $B = B$, $H = \mathscr{L}^2$, $R = b\|\tilde{L}\|_{B \to B}$, $S(v, y) = v - K_2^{\frac{1}{2}} M_1(u_0(y) + K_1^{\frac{1}{2}} v)$, $T(v, y) = \tilde{L} M_2(u_0(y) + K_1^{\frac{1}{2}} v)$. Die Überprüfung der Voraussetzungen des Hilfssatzes 1 verläuft ganz analog wie beim Beweis von Satz 1.

Um Existenz- und Eindeutigkeitssätze für „verallgemeinerte" Lösungen von (3.1) aufzustellen, benötigt man weitere Voraussetzungen.

Voraussetzung 3. Außer Voraussetzung 1 gelte:

(IX) Es existieren Konstanten $R_0 > 0$ und $M_{R_0} > 0$, so daß für $y \in B$ mit $\|y\|_B \leqq R_0$ gelte

$$\|M_1(u_0(y))\|_{\mathscr{L}^q} \leqq M_{R_0} / \|\mathscr{H}^*\|_{\mathscr{L}^q \to \mathscr{L}^2}.$$

(X) Es existiere eine Konstante $L_{R_0}>0$, so daß für $v\in\mathscr{L}^2$, $y_j\in B$ $(j=1,2)$ mit $\|v\|_{\mathscr{L}^2}\leqq M_{R_0}/C$, $\|y_j\|_B\leqq R_0$ gelte

$$\|M_1(\mathscr{H}v+u_0(y_1))-M_1(\mathscr{H}v+u_0(y_2))\|_{\mathscr{L}^q}\leqq L_{R_0}\|u_0(y_1)-u_0(y_2)\|_{\mathscr{L}^p}/(\|\mathscr{H}^*\|_{\mathscr{L}^q\to\mathscr{L}^2}\|u_0\|_{B\to\mathscr{L}^p}).$$

(XI) Es existiere eine Konstante $k>0$, so daß für $v_j\in\mathscr{L}^2$, $y_j\in B$ $(j=1,2)$ mit $\|v_j\|_{\mathscr{L}^2}\leqq M_{R_0}/C$, $\|y_j\|_B\leqq R_0$ gelte

$$\|M_2(\mathscr{H}v_1+u_0(y_1))-M_2(\mathscr{H}v_2+u_0(y_2))\|_B\leqq k/(\|\mathscr{H}\|_{\mathscr{L}^2\to\mathscr{L}^p}\|\tilde{L}\|_{B\to B})\,\|\mathscr{H}(v_1-v_2)+u_0(y_1-y_2)\|_{\mathscr{L}^p}$$

mit

$$k(L_{R_0}/C+\|u_0\|_{B\to\mathscr{L}^p}/\|\mathscr{H}\|_{\mathscr{L}^2\to\mathscr{L}^p})<1.$$

(XII) Aus $Kz=\Theta_{\mathscr{L}^p}$ mit $z\in\mathscr{L}^q$ folge $z=\Theta_{\mathscr{L}^q}$.

Voraussetzung 4. Außer Voraussetzung 2 gelte:

(IX′) Es existieren Konstanten $R_0>0$ und $M_{R_0}>0$, so daß für $y\in B$ mit $\|y\|_B\leqq R_0$ gelte

$$\|M_1(u_0(y))\|_{B_2}\leqq M_{R_0}/\|K_2^{\frac{1}{2}}\|_{B_2\to\mathscr{L}^2}.$$

(X′) Es existiere eine Konstante $L_{R_0}>0$, so daß für $v\in\mathscr{L}^2$, $y_j\in B$ $(j=1,2)$ mit $\|v\|_{\mathscr{L}^2}\leqq M_{R_0}/C$, $\|y_j\|_B\leqq R_0$ gelte

$$\|M_1(K_1^{\frac{1}{2}}v+u_0(y_1))-M_1(K_1^{\frac{1}{2}}v+u_0(y_2))\|_{B_2}\leqq L_{R_0}\|u_0(y_1)-u_0(y_2)\|_{B_1}/(\|K_2^{\frac{1}{2}}\|_{B_2\to\mathscr{L}^2}\|u_0\|_{B\to B_1}).$$

(XI′) Es existiere eine Konstante $k>0$, so daß für $v_j\in\mathscr{L}^2$, $y_j\in B$ $(j=1,2)$ mit $\|v_j\|_{\mathscr{L}^2}\leqq M_{R_0}/C$, $\|y_j\|_B\leqq R_0$ gelte

$$\|M_2(K_1^{\frac{1}{2}}v_1+u_0(y_1))-M_2(K_1^{\frac{1}{2}}v_2+u_0(y_2))\|_B\leqq k/(\|K_1^{\frac{1}{2}}\|_{\mathscr{L}^2\to B_1}\|\tilde{L}\|_{B\to B})\,\|K_1^{\frac{1}{2}}(v_1-v_2)+u_0(y_1-y_2)\|_{B_1}$$

mit

$$k(L_{R_0}/C+\|u_0\|_{B\to B_1}/\|K_1^{\frac{1}{2}}\|_{\mathscr{L}^2\to B_1})<1.$$

(XII′) Aus $Kz=\Theta_{B_1}$ mit $z\in B_2$ folge $z=\Theta_{B_2}$.

Es gilt der folgende lokale Existenz- und Pseudo-Eindeutigkeitssatz:

Satz 3. Es sei die Voraussetzung 3 erfüllt. Dann gibt es genau eine „verallgemeinerte" Lösung

$$u^*:=u_0(y^*)+\mathscr{H}v^*\in\mathscr{L}^p \tag{3.21}$$

von (3.1) mit $y^*\in B$, $v^*\in\mathscr{L}^2$, für welche $\|y^*\|_B\leqq R_0$ gilt. Ferner konvergiert das folgende Iterationsverfahren

$$\begin{aligned} v_{\nu+1}&=\mathscr{H}^*M_1(\mathscr{H}v_{\nu+1}+u_0(y_\nu))\\ y_{\nu+1}&=\tilde{L}M_2(\mathscr{H}v_{\nu+1}+u_0(y_\nu))\quad(\nu=0,1,2\ldots)\end{aligned} \tag{3.22}$$

mit $y_0\in B$ und $\|y_0\|_B\leqq R_0$ gegen die Elemente v^*, y^*.

Beweis. Wir betrachten zunächst die Gln. (3.14) und beweisen die Existenz von genau einer Lösung v^*, y^* mit $\|y^*\|_B \leqq R_0$ und die Konvergenz des Iterationsverfahrens (3.22). Dies folgt durch Anwendung von Hilfssatz 2. Man setze mit den Bezeichnungen des Hilfssatzes:

$$S(v, y) = v - \mathscr{H}^* M_1(\mathscr{H} v + u_0(y)), \qquad T(v, y) = \tilde{L} M_2(\mathscr{H} v + u_0(y)).$$

Die Bedingungen (A'α, β), (B') und (D') folgen analog wie beim Beweis von Satz 1. Überprüfung der restlichen Bedingungen des Hilfssatzes:

(A'γ) Sei $v \in \mathscr{L}^2$, $y_j \in B$ $(j = 1, 2)$ mit $\|v\|_{\mathscr{L}^2} \leqq M_{R_0}/C$ und $\|y_j\|_B \leqq R_0$, dann folgt mittels (X)

$$\begin{aligned}\|S_0(v, y_1) - S_0(v, y_2)\|_{\mathscr{L}^2} &= \|\mathscr{H}^*(M_1(\mathscr{H} v + u_0(y_1)) - M_1(\mathscr{H} v + u_0(y_2)))\|_{\mathscr{L}^2} \\ &\leqq \|\mathscr{H}^*\|_{\mathscr{L}^q \to \mathscr{L}^2} L_{R_0} \|u_0(y_1 - y_2)\|_{\mathscr{L}^p} / (\|\mathscr{H}^*\|_{\mathscr{L}^q \to \mathscr{L}^2} \|u_0\|_{B \to \mathscr{L}^p}) \\ &\leqq L_{R_0} \|y_1 - y_2\|_B .\end{aligned}$$

(C') Sei $y \in B$ mit $\|y\|_B \leqq R_0$, dann gilt nach (IX)

$$\|S(\Theta_{\mathscr{L}^2}, y)\|_{\mathscr{L}^2} = \|\mathscr{H}^* M_1(u_0(y))\|_{\mathscr{L}^2} \leqq M_{R_0}.$$

(E') Sei $v_j \in \mathscr{L}^2$, $y_j \in B$ $(j = 1, 2)$ mit $\|v_j\|_{\mathscr{L}^2} \leqq M_{R_0}/C$, $\|y_j\|_B \leqq R_0$, dann erhält man mittels (XI)

$$\begin{aligned}\|T(v_1, y_1) - T(v_2, y_2)\|_B &\leqq \frac{k}{\|\mathscr{H}\|_{\mathscr{L}^2 \to \mathscr{L}^p}} \|\mathscr{H}(v_1 - v_2) + u_0(y_1 - y_2)\|_{\mathscr{L}^p} \\ &\leqq k \|v_1 - v_2\|_{\mathscr{L}^2} + (k \|u_0\|_{B \to \mathscr{L}^p} / \|\mathscr{H}\|_{\mathscr{L}^2 \to \mathscr{L}^p}) \|y_1 - y_2\|_B ,\end{aligned}$$

womit mit

$$k_1 = k, \qquad k_2 = k \|u_0\|_{B \to \mathscr{L}^p} / \|\mathscr{H}\|_{\mathscr{L}^2 \to \mathscr{L}^p}$$

Bedingung (E') erfüllt ist.

Es existiert daher nach Hilfssatz 2 genau eine Lösung v^*, y^* von (3.14) mit $\|y^*\|_B \leqq R_0$, welche nach (3.22) konstruiert werden kann. Nach Lemma 1 folgt hieraus die Existenz einer „verallgemeinerten" Lösung u^*, die in der Form (3.21) mit $\|y^*\|_B \leqq R_0$ darstellbar ist.

Wir zeigen nun, daß es genau eine „verallgemeinerte" Lösung $u^* \in \mathscr{L}^p$ von (3.1) mit $\|y^*\|_B \leqq R_0$ gibt. Dazu sei angenommen, daß u_j^*, y_j^* $(j = 1, 2)$ mit $\|y_j^*\|_B \leqq R_0$ zwei Lösungen von (3.16) seien, d. h. es gilt wegen (V)

$$\begin{aligned} u_j^* &= u_0(y_j^*) + K M_1(u_j^*) \\ y_j^* &= \tilde{L} M_2 u_j^* . \end{aligned} \tag{2.23}$$

Aus der ersten dieser Beziehungen folgt für u_j^* die Darstellung

$$u_j^* = u_0(y_j^*) + K z_j^* \tag{2.24}$$

mit $z_j^* \in \mathscr{L}^q$. Nach (XII) folgt aus (2.23) und (2.24) die Beziehung

$$z_j^* = M_1(u_j^*). \tag{2.25}$$

Wendet man hierauf den Operator $\mathscr{H}^*$ an und setzt

$$v_j^* = \mathscr{H}^* z_j^* ,$$

so erhält man wegen (2.24) und (III)

$$v_j^* = \mathscr{H}^* M_1\left(\mathscr{H} v_j^* + u_0(y_j^*)\right)$$
$$y_j^* = \tilde{L} M_2\left(\mathscr{H} v_j^* + u_0(y_j^*)\right).$$

v_j^*, y_j^* $(j=1, 2)$ ist also Lösung von (3.14), und nach dem ersten Teil des Beweises existiert genau ein v^*, y^* von (3.14) mit $\|y^*\|_B \leqq R_0$. Es gilt daher $v^* = v_j^*$, $y^* = y_j^*$ und somit $u^* = u_j^*$ $(j=1, 2)$.

Damit ist Satz 3 bewiesen.

Ein analoger Satz gilt unter der Voraussetzung 4.

Satz 4. Es sei die Voraussetzung 4 erfüllt. Dann gibt es genau eine „verallgemeinerte" Lösung

$$u^* = u_0(y^*) + K_1^{\frac{1}{2}} v^* \in B_1 \tag{3.26}$$

von (3.1) mit $y^* \in B$, $v^* \in \mathscr{L}^2$, für welche $\|y^*\|_B \leqq R_0$ gilt. Ferner konvergiert das folgende Iterationsverfahren

$$\begin{aligned} v_{\nu+1} &= K_2^{\frac{1}{2}} M_1\left(K_1^{\frac{1}{2}} v_{\nu+1} + u_0(y_\nu)\right) \\ y_{\nu+1} &= \tilde{L} M_2\left(K_1^{\frac{1}{2}} v_{\nu+1} + u_0(y_\nu)\right) \end{aligned} \tag{3.27}$$

mit $y_0 \in B$ und $\|y_0\|_B \leqq R_0$ gegen die Elemente v^*, y^*.

Beweis. Man betrachtet zunächst die Gln. (3.17). Der Beweis verläuft dann analog wie derjenige von Satz 3.

Bemerkung 3. Sind die Bedingungen (IX)–(XI) von Voraussetzung 3 bzw. (IX')–(XI') von Voraussetzung 4 für alle $R_0 \geqq R^*$ mit einem geeigneten $R^* > 0$ erfüllbar, dann besitzt das Problem (3.1) genau eine „verallgemeinerte" Lösung.

Damit diese Voraussetzung erfüllt sein kann, muß notwendigerweise eine der beiden folgenden Bedingungen gelten:

(a) Es ist $k=0$, d.h. es existiert $f \in B$, so daß für alle $u \in \mathscr{L}^p$ bzw. $u \in B_1$ gilt $M_2 u = f$ (lineare Randbedingung).

(b) Es existiert eine Konstante $L^* > 0$, so daß für alle $R_0 \geqq R^*$ gilt $L_{R_0} \leqq L^*$, d.h. M_1 genügt einer globalen Lipschitzbedingung.

Beweis. Die Behauptung folgt unmittelbar aus Satz 3 bzw. 4 unter Beachtung von Voraussetzung (XI) bzw. (XI').

4. Anwendungen

In diesem Abschnitt sollen zwei Anwendungen der allgemeinen Existenzsätze gebracht werden.

Beispiel 1. Als erste Anwendung betrachten wir das nichtlineare Randwertproblem

$$\begin{aligned} -\Delta u(s) &= f\left(s, u(s)\right) \quad \text{für} \quad s \in \Omega \\ u(s) &= \left(M_2 u\right)(s) \quad \text{für} \quad s \in \partial\Omega. \end{aligned} \tag{4.1}$$

Hierbei bezeichne Ω eine offene beschränkte konvexe Menge aus R^2, $\partial\Omega$ den Rand von Ω und Δ den zweidimensionalen Laplace-Operator.

Wir machen nun die folgende Voraussetzung:

(a) Ω sei eine offene beschränkte konvexe Menge aus R^2, und der Rand $\partial\Omega$ sei doppelpunktfrei und stetig gekrümmt[5].

Bemerkung 4 (vgl. z.B. [9], S. 176ff.). Unter der Voraussetzung (a) gilt für $\mu \in C^0[\partial\Omega]$ mit der Abkürzung

$$W_\mu(s) = \frac{1}{\pi} \int_{\partial\Omega} \frac{\cos(|s,t|, n_t)}{|s,t|} \mu(t)\, dt \tag{4.2}$$

[6]

die Relation

$$\Delta W_\mu(s) \equiv 0 \quad \text{für} \quad s \in \Omega, \tag{4.3}$$

und für $s_k \in \Omega$, $s \in \partial\Omega$ und $s_k \to s$ gilt

$$\lim_{s_k \to s} W_\mu(s_k) = \mu(s) + W_\mu(s). \tag{4.4}$$

Ferner ist die Funktion

$$A(s,t) = \frac{1}{\pi} \frac{\cos(|s,t|, n_t)}{|s,t|}$$

stetig auf $\partial\Omega \times \partial\Omega$ und das Problem

$$\mu(s) + \int_{\partial\Omega} A(s,t)\, \mu(t)\, dt = 0$$

mit $\mu \in C^0[\partial\Omega]$ besitzt nur die triviale Lösung $\mu(t) \equiv 0$ auf $\partial\Omega$.

Bemerkung 5 (vgl. z.B. [2] Bd. I, S. 316 und Bd. II, S. 282ff.). Unter der Voraussetzung (a) gibt es eine Funktion (Greensche Funktion)

$$K(s,t) = \frac{1}{2\pi} \log \frac{1}{|s,t|} + K_0(s,t) \tag{4.5}$$

mit $K_0 \in C^0[\bar{\Omega} \times \bar{\Omega}]$, $\Delta_s K_0(s,t) = 0$ für $s, t \in \Omega$, $K(s,t) = 0$ für $s \in \partial\Omega$, $t \in \Omega$ und $K(s,t) = K(t,s)$ für $s \neq t$ und $s, t \in \bar{\Omega}$. Für $f \in C^1[\bar{\Omega}]$ folgen aus

$$u(s) = \int_\Omega K(s,t)\, f(t)\, dt \tag{4.6}$$

die Beziehungen

$$\begin{aligned} -\Delta u(s) &= f(s) \quad \text{für} \quad s \in \Omega \\ u(s) &= 0 \quad \text{für} \quad s \in \partial\Omega. \end{aligned} \tag{4.7}$$

[7]

Ferner ist die Greensche Funktion $K(s,t)$ positiv definit (vgl. z.B. [2] Bd. I, S. 320).

Es seien ferner die beiden folgenden Voraussetzungen erfüllt:

(b) $f(s,u)$ sei eine in $\mathfrak{H} = \{(s,u) \mid s \in \bar{\Omega}, u \in R^1\}$ stetige reelle Funktion mit

$$(f(s,u_1) - f(s,u_2))(u_1 - u_2) \leqq a(u_1 - u_2)^2 \tag{4.8}$$

5 Diese Bedingung kann abgeschwächt werden.

6 $|s,t|$ bedeutet hierbei den euklidischen Abstand zwischen s und t in R^2 und n_t die innere Normale an $\partial\Omega$ im Punkte t.

7 $u(s) = 0$ für $s \in \partial\Omega$ bedeutet hierbei $u(s_k) \to 0$ für $s_k \in \Omega$ und $s_k \to s \in \partial\Omega$.

für $s \in \bar{\Omega}$, $u_j \in R^1$ $(j = 1, 2)$ und mit

$$a \geqq 0, \quad C := 1 - a/\lambda_1 > 0. \tag{4.9}$$

Hierbei sei λ_1 der kleinste Eigenwert von

$$\varphi(s) = \lambda \int_{\Omega} K(s, t)\, \varphi(t)\, dt$$

mit $\varphi \in C^0[\bar{\Omega}]$.

Ferner existieren Konstanten $D_j \geqq 0$ $(j = 1, 2)$ und $2 \leqq p < \infty$, so daß für $s \in \bar{\Omega}$, $u \in R^1$ gelte

$$|f(s, u)| \leqq D_1 + D_2 |u|^{p-1}. \tag{4.10}$$

(c) M_2 sei eine (nichtlineare) vollstetige Abbildung von $L^p[\Omega]$ in $C^0[\partial\Omega]$. Es existiere eine Konstante $b > 0$, so daß für alle $u \in L^p[\Omega]$ gelte

$$\|M_2 u\|_{C^0[\partial\Omega]} \leqq b. \tag{4.11}$$

Wir beweisen nun als Anwendung von Satz 1 das folgende

Lemma 4. Es seien die Voraussetzungen (a), (b) und (c) erfüllt. Dann besitzt das Problem (4.1) mindestens eine „verallgemeinerte" Lösung.

Beweis. Wir wenden Satz 1 an. Mit den Bezeichnungen des Satzes setzen wir:

$$B := C^0[\partial\Omega], \quad \mathscr{L}^p := L^p[\Omega], \quad \mathscr{L}^q := L^q[\Omega] \left(q = \frac{p}{p-1}\right). \tag{4.12a}$$

$$\begin{aligned} &(L_1 u)(s) := -\Delta u(s), \quad (M_1 u)(s) := f(s, u(s)) \quad (s \in \Omega), \\ &(L_2 u)(s) := \lim_{s_k \to s} u(s_k) \quad \text{für} \quad s_k \in \Omega,\ s \in \partial\Omega \end{aligned} \tag{4.12b}$$

mit

$$\begin{aligned} D(L_1) &:= C^2[\Omega] \cap C^0[\bar{\Omega}], \quad D(L_2) := C^0[\bar{\Omega}], \\ \tilde{B} &:= C^0[\Omega], \qquad\qquad \tilde{B}_0 := C^1[\bar{\Omega}]. \end{aligned} \tag{4.12c}$$

$$u_0(y)(s) := W_y(s) \qquad (s \in \Omega). \tag{4.12d}$$

$$(K u)(s) := \int_{\Omega} K(s, t)\, u(t)\, dt \qquad (s \in \Omega). \tag{4.12e}$$

Mit diesen Festlegungen überprüfen wir die Voraussetzungen von Satz 1.

(I) Nach (4.12a, b, c) ist wegen $|\Omega| < \infty$ L_1 ein linearer Operator mit $D(L_1) \subset \mathscr{L}^p$ und $R(L_1) \subset C^0[\Omega] = \tilde{B}$. Ebenso ist L_2 ein linearer Operator mit $D(L_1) \subset D(L_2) \subset \mathscr{L}^p$ und $R(L_2) \subset C^0[\partial\Omega] = B$. M_1 ist nach (12b) und Voraussetzung (4.10) eine stetige beschränkte Abbildung von $L^p[\Omega]$ in $L^q[\Omega]$ (vgl. z.B. [10], S. 154 Satz 19.1 Nr. 2) und M_2 ist nach Voraussetzung (c) eine stetige beschränkte Abbildung von $L^p[\Omega]$ in $C^0[\partial\Omega]$.

(II) Der erste und letzte Teil der Voraussetzung folgen nach (4.12e) aus Bemerkung 5. Es ist noch zu zeigen, daß K ein linearer beschränkter Operator von $L^q[\Omega]$ in $L^p[\Omega]$ ist. Dies wird zusammen mit Voraussetzung (III) bewiesen.

(III) Es genügt wegen $|\Omega| < \infty$ mit einem $\varepsilon > 0$ zu zeigen

$$\int_\Omega \int_\Omega |K(s,t)|^{p+\varepsilon}\, dt\, ds < \infty$$

(vgl. z.B. [7], S. 56 und 52), was unmittelbar aus (4.5) folgt. Hiermit ist Voraussetzung (III) und der Rest von Voraussetzung (II) bewiesen.

(IV) Bedingung (3.4) folgt aus (4.3), (4.4) und (4.12d). Für $y \in C^0[\partial\Omega]$ gilt nach (4.12)

$$\begin{aligned}\|u_0(y)\|_{\mathscr{L}^p} &= \Big(\int_\Omega \Big|\int_{\partial\Omega} A(s,t)\, y(t)\, dt\Big|^p ds\Big)^{1/p} \\ &\leqq \Big(\int_\Omega \Big(\int_{\partial\Omega} |A(s,t)|\, dt\Big)^p ds\Big)^{1/p} \|y\|_{C^0[\partial\Omega]}.\end{aligned}$$

Da nach Voraussetzung (a) die Menge Ω konvex ist[8], gilt für $s \in \bar{\Omega}$, $t \in \partial\Omega$

$$A(s,t) = \frac{1}{\pi} \frac{\cos(|s,t|, n_t)}{|s,t|} \geqq 0,$$

und wegen (4.4) und

$$\int_{\partial\Omega} A(s,t)\, dt = 1 \quad \text{für} \quad s \in \partial\Omega$$

(vgl. z.B. [6], S. 130) erhält man

$$\int_{\partial\Omega} A(s,t)\, dt = 2 \quad \text{für} \quad s \in \Omega.$$

Zusammenfassend ergibt dies die Abschätzung

$$\begin{aligned}\|u_0(y)\|_{\mathscr{L}^p} &\leqq \Big(\int_\Omega \Big(\int_{\partial\Omega} A(s,t)\, dt\Big)^p ds\Big)^{1/p} \|y\|_{C^0[\partial\Omega]} \\ &\leqq 2\,|\Omega|^{1/p} \|y\|_{C^0[\partial\Omega]},\end{aligned}$$

womit Bedingung (IV) bewiesen ist.

(V) Zunächst folgt mit $y \in C^0[\partial\Omega]$ aus (4.4) und (4.12b, d) für $s \in \partial\Omega$

$$\big(L_2 u_0(y)\big)(s) = y(s) + \int_{\partial\Omega} A(s,t)\, y(t)\, dt =: (I + A)\, y.$$

Nun ist nach Bemerkung 4 die Abbildung A vollstetig auf $C^0[\partial\Omega]$ und -1 ist kein Eigenwert. Nach dem Alternativsatz für vollstetige Operatoren existiert daher eine beschränkte Inverse $\tilde{L} := (I + A)^{-1}$ auf $C^0[\partial\Omega]$, d.h.

$$\|\tilde{L}\|_{B\to B} = \|(I + A)^{-1}\|_{B\to B} < \infty,$$

womit Bedingung (V) bewiesen ist.

(VI) Sei $u_j \in L^p[\Omega]$ $(j = 1, 2)$, dann folgt wegen $p \geqq 2$ und $f(s, u_j(s)) \in L^q[\Omega]$ (vgl. (I)) aus (4.8)

$$\begin{aligned}&\int_\Omega \big\langle \big(M_1(u_1) - M_1(u_2)\big)(s), (u_1 - u_2)(s)\big\rangle\, ds \\ &\qquad = \int_\Omega \big(f(s, u_1(s)) - f(s, u_2(s))\big)\big(u_1(s) - u_2(s)\big)\, ds \leqq a\, \|u_1 - u_2\|^2_{\mathscr{L}^2}.\end{aligned}$$

8 Die Voraussetzung, daß Ω konvex ist kann abgeschwächt werden.

Die zweite Bedingung von (3.7) folgt aus (4.9) unter Beachtung von

$$\|\mathscr{H}\|_{L^2[\Omega]\to L^2[\Omega]} = \frac{1}{\sqrt{\lambda_1}}.$$

(VII) Die Bedingung folgt aus (4.11).

(VIII) Da nach Voraussetzung (c) M_2 vollstetig ist von $L^p[\Omega]$ in $C^0[\partial\Omega]$ und $\tilde{L}$ beschränkt auf $C^0[\partial\Omega]$ (vgl. (V)), ist $\tilde{L}M_2$ kompakt von $L^p[\Omega]$ in $C^0[\partial\Omega]$. Damit sind die Voraussetzungen von Satz 1 erfüllt. Die Behauptung von Lemma 4 folgt aus Satz 1.

Nach Lemma 4 existieren somit Funktionen $y^*\in C^0[\partial\Omega]$ und u^* mit $\int\limits_\Omega |u^*(s)|^p\,ds<\infty$, so daß gilt

$$u^*(s) = \int_{\partial\Omega} A(s,t)\,y^*(t)\,dt + \int_\Omega K(s,t)\,f\big(t,u^*(t)\big)\,dt \tag{4.13a}$$

fast überall auf Ω und

$$y^*(s) + \int_{\partial\Omega} A(s,t)\,y^*(t)\,dt = M_2(u^*)(s) \tag{4.13b}$$

identisch für $s\in\partial\Omega$.

Um die Existenz einer Lösung von (4.1) im „klassischen" Sinn zu sichern machen wir die folgende zusätzliche Voraussetzung:

(d) Es seien $\frac{\partial}{\partial s_i} f(s,u)$ $(i=1,2)$ und $\frac{\partial}{\partial u} f(s,u)$ stetige Funktionen in $\mathfrak{H}$. Für $u\in L^p[\Omega]$ gelte $M_2 u\in C^\alpha[\partial\Omega]$ mit $\alpha>0$ (Klasse der α-hölderstetigen Funktionen).

Dann gilt der

Satz 5. Es seien die Voraussetzungen (a), (b), (c) und (d) erfüllt. Dann gibt es in der Äquivalenzklasse der Lösungsfunktionen aus $L^p[\Omega]$ von (4.13) eine Funktion $u^*\in C^2[\Omega]\cap C^\delta[\bar{\Omega}]$ mit $\delta>0$, welche das Problem (4.1) im „klassischen" Sinn löst.

Beweis. Für $s_1, s_2\in\bar{\Omega}$ folgt wegen $f\big(t,u^*(t)\big)\in L^q[\Omega]$

$$\begin{aligned}\Big|\int_\Omega K(s_1,t)\,f\big(t,u^*(t)\big)\,dt - \int_\Omega K(s_2,t)\,f\big(t,u^*(t)\big)\,dt\Big| \\ \leqq \Big(\int_\Omega |K(s_1,t)-K(s_2,t)|^p\,dt\Big)^{1/p}\Big(\int_\Omega |f\big(t,u^*(t)\big)|^q\,dt\Big)^{1/q}\to 0\end{aligned}$$

für $s_1\to s_2$, d.h. es gilt

$$\int_\Omega K(s,t)\,f\big(t,u^*(t)\big)\,dt\in C^0[\bar{\Omega}]. \tag{4.14}$$

Nach (4.13a) kann daher wegen (4.3), (4.4), (4.12d) und (4.14) aus der Äquivalenzklasse der Lösungsfunktionen von (4.13) $u^*\in C^0[\bar{\Omega}]$ gewählt werden, und mit diesem u^* ist die Beziehung (4.13a) identisch auf Ω erfüllt. Aus (4.13a) folgt daher mittels (4.4), (4.14) und Bemerkung 5 für $s\in\Omega$ und $s\to\bar{s}\in\partial\Omega$

$$\lim_{s\to\bar{s}} u^*(s) = y^*(\bar{s}) + \int_{\partial\Omega} A(\bar{s},t)\,y^*(t)\,dt,$$

d.h. zusammen mit (4.13b) erhält man

$$u^*(s) = M_2(u^*)(s)$$

für $s \in \partial\Omega$, womit die zweite Beziehung von (4.1) bewiesen ist. Wegen $u^* \in C^0[\bar{\Omega}]$ ist $f(s, u^*(s)) \in C^0[\bar{\Omega}]$, d.h. es gilt (vgl. z.B. [2], Bd. I, S. 317)

$$\int_\Omega K(s,t) f(t, u^*(t))\, dt \in C^1[\bar{\Omega}]. \tag{4.15}$$

Ebenso ist wegen $y^* \in C^0[\partial\Omega]$ (vgl. z.B. [5], S. 47)

$$\int_{\partial\Omega} A(s,t)\, y^*(t)\, dt \in C^\varepsilon[\partial\Omega]$$

mit $\varepsilon > 0$ und somit folgt wegen $M_2(u^*) \in C^\alpha[\partial\Omega]$ (Voraussetzung (d)) aus (4.13b) $y^* \in C^\delta[\partial\Omega]$ mit $\delta > 0$. Hieraus ergibt sich (vgl. [5], S. 48)

$$\int_{\partial\Omega} A(s,t)\, y^*(t)\, dt \in C^\delta[\bar{\Omega}],$$

und nach (4.13a) und (4.15) gilt daher $u^* \in C^\delta[\bar{\Omega}]$. Nach Voraussetzung (d) ist somit $f(s, u^*(s)) \in C^\delta[\bar{\Omega}]$ und nach dem Satz von Poisson erhält man schließlich (vgl. [5], S. 87) $\int_\Omega K(s,t) f(t, u^*(t))\, dt \in C^2[\Omega]$ mit

$$-\Delta \int_\Omega K(s,t) f(t, u^*(t))\, dt = f(s, u^*(s)) \quad \text{für} \quad s \in \Omega.$$

Aus (4.13a) folgt daher mittels (4.3) die erste Beziehung von (4.1). Damit ist Satz 5 bewiesen.

Bemerkung 6a. Es seien $k(s,t)$, $\frac{\partial}{\partial s} k(s,t) \in C^0[\partial\Omega \times \bar{\Omega}]$ und $a(s)$, $b(s) \in C^1[\partial\Omega]$. Für $u \in L^p[\Omega]$ wird beispielsweise durch

$$(M_2 u)(s) := a(s) + b(s) \Big/ \Big(1 + \Big(\int_\Omega k(s,t)\, u(t)\, dt\Big)^2\Big) \quad (s \in \partial\Omega)$$

eine Abbildung M_2 erzeugt, die den Voraussetzungen (c) und (d) genügt.

Bemerkung 6b. Es sei $c(s)$, $d(s) \in C^1[\bar{\Omega}]$ mit $d(s) \geqq 0$ auf $\bar{\Omega}$. Ferner sei

$$f(s,u) = c(s) - d(s)\, u^{2m-1}/(1+u^2)$$

mit $m \geqq 2$ und ganzzahlig. Mit $p = 2m$, $a = 0$ und $C = 1$ sind dann die Voraussetzungen (b) und (d) erfüllt.

Beweis. Die Bedingung (4.8) folgt mittels des Mittelwertsatzes der Differentialrechnung. Der Rest ist trivial.

Als nächste Anwendung betrachten wir das

Beispiel 2. Vorgelegt sei die nichtlineare Randwertaufgabe

$$\begin{aligned} -u''(s) &= g(s, u(s)) \quad \text{für} \quad s \in \Omega :=]0,1[, \\ u(0) &= F_1(u), \qquad u(1) = F_2(u). \end{aligned} \tag{4.16}$$

Wir machen nun die folgenden Voraussetzungen:

(a') $g(s, u)$ sei eine in $\mathfrak{H}_0 = \{(s, u) \mid s \in \Omega,\ u \in R^1\}$ stetige reelle Funktion mit

$$(g(s, u_1) - g(s, u_2))(u_1 - u_2) \leqq a(u_1 - u_2)^2 \tag{4.17}$$

für $s \in \Omega$, $u_j \in R^1$ $(j = 1, 2)$ und mit

$$a \geqq 0, \qquad C := 1 - \frac{a}{\pi^2} > 0. \tag{4.18}$$

Ferner existieren auf Ω stetige Funktionen $g_j(s)$ $(j = 1, 2)$ mit $\int_\Omega (g_j(s))^2\, ds < \infty$ und eine auf $R^1_+ = \{u \mid u \in R^1,\ u \geqq 0\}$ stetige Funktion $h(u)$, so daß für $s \in \Omega$ und $u \in R^1$ gelte

$$|g(s, u)| \leqq g_1(s) + g_2(s)\, h(|u|). \tag{4.19}$$

(b') Es seien F_j $(j = 1, 2)$ stetige Abbildungen von $C^0[\bar{\Omega}]$ in R^1 mit

$$|F_j(u)| \leqq b \quad (j = 1, 2) \tag{4.20}$$

für alle $u \in C^0[\bar{\Omega}]$ mit einer geeigneten Konstanten $b > 0$.

Als Anwendung von Satz 2 beweisen wir das

Lemma 5. Die Voraussetzungen (a') und (b') seien erfüllt. Dann besitzt das Problem (4.16) mindestens eine „verallgemeinerte" Lösung.

Beweis. Wir wenden Satz 2 an. Mit den Bezeichnungen des Satzes setzen wir:

$$\begin{gathered} \mathscr{L}^2 = B_2 = L^2[\Omega], \qquad B_1 = C_0[\bar{\Omega}], \qquad B = R^2 \\ \text{mit der Norm} \quad \|y\|_B = \operatorname{Max}(|y_1|, |y_2|) \quad \text{für} \quad y = \{y_1, y_2\} \in B, \end{gathered} \tag{4.21a}$$

$$\begin{gathered} (L_1 u)(s) = -u''(s), \qquad (M_1 u)(s) = g(s, u(s)) \quad (s \in \Omega), \\ L_2 u = \{u(0), u(1)\}, \qquad M_2 u = \{F_1 u, F_2 u\} \end{gathered} \tag{4.21b}$$

mit

$$\begin{gathered} D(L_1) = C^2[\Omega] \cap C^0[\bar{\Omega}], \qquad D(L_2) = C^0[\bar{\Omega}], \\ \tilde{B} = C^0[\Omega], \qquad \tilde{B}_0 = C^0[\Omega] \cap L^2[\Omega], \end{gathered} \tag{4.21c}$$

$$u_0(y)(s) := y_1 + y_2 s \qquad \text{für} \quad s \in \bar{\Omega} \quad \text{und} \quad y = \{y_1, y_2\} \in R^2, \tag{4.21d}$$

$$(K u)(s) = \int_0^1 K(s, t)\, u(t)\, dt \quad \text{für} \quad s \in \bar{\Omega} \tag{4.21e}$$

mit

$$K(s, t) = \begin{cases} (1 - s)\, t, & 0 \leqq t \leqq s \\ s(1 - t), & s \leqq t \leqq 1. \end{cases} \tag{4.21f}$$

Mit diesen Festlegungen überprüfen wir die Voraussetzungen von Satz 2.

(I') L_1 ist nach (4.21a, b, c) ein linearer Operator mit $D(L_1) \subset B_1 \subset L^2[\Omega]$ und $R(L_1) \subset C^0[\Omega] = \tilde{B}$. Ebenso ist L_2 nach (4.21a, b, c) ein linearer Operator mit $D(L_1) \subset D(L_2) = B_1$ und $R(L_2) \subset B$. Da nach Voraussetzung (a') $g(s, u)$ stetig ist in $\mathfrak{H}_0$ und Bedingung (4.19) gilt, ist M_1 nach dem Satz von Lebesgue stetig und beschränkt von $C^0[\bar{\Omega}]$ in $L^2[\Omega]$. Nach Voraussetzung (b') ist M_2 ein stetiger beschränkter Operator von $C^0[\bar{\Omega}]$ in B.

(II′) Sei $f \in \tilde{B}_0$, und für $s \in \bar{\Omega}$ gelte

$$u(s) = \int_{\Omega} K(s,t)\, f(t)\, dt,$$

dann folgt durch elementare Rechnung

$$-u''(s) = f(s) \quad (s \in \Omega), \qquad u(0) = u(1) = 0.$$

Sei $v \in L^2[\Omega]$, dann gilt für $s \in \bar{\Omega}$ die Darstellung

$$(Kv)(s) := \sum_{i=1}^{\infty} \frac{(\varphi_i, v)}{\lambda_i}\, \varphi_i(s) \tag{4.22a}$$

mit

$$\varphi_i(s) = \sqrt{2} \sin \pi i s, \qquad \lambda_i = i^2 \pi^2. \tag{4.22b}$$

Der durch (4.21e, f) erzeugte Operator K ist linear und beschränkt von $L^2[\Omega]$ in $C^0[\bar{\Omega}]$. Ebenso wird nach (4.22) durch $K(s,t)$ ein linearer, beschränkter, selbstadjungierter, positiv definiter Operator $\tilde{K}$ auf $L^2[\Omega]$ erzeugt, für welchen die Beziehung $J_1 K = \tilde{K}$ besteht, (J_1 ist die Injektionsabbildung von $C^0[\bar{\Omega}]$ in $L^2[\Omega]$).

(III′) Da J_2 die identische Abbildung ist, gilt $K_2^{\frac{1}{2}} = K^{\frac{1}{2}}$ (eindeutig bestimmter positiv definiter Wurzeloperator von $\tilde{K}$). Für $v \in L^2[\Omega]$ setzen wir mit $s \in \bar{\Omega}$

$$(K_1^{\frac{1}{2}} v)(s) := \sum_{i=1}^{\infty} \frac{(\varphi_i, v)}{\sqrt{\lambda_i}}\, \varphi_i(s).$$

Es bleibt noch zu zeigen, daß $K_1^{\frac{1}{2}}$ linear und beschränkt von $L^2[\Omega]$ in $C^0[\bar{\Omega}]$ ist. Mittels der Schwarzschen Ungleichung erhält man

$$\begin{aligned} \operatorname*{Max}_{s, s+h \in \bar{\Omega}} \left| (K_1^{\frac{1}{2}} v)(s+h) - (K_1^{\frac{1}{2}} v)(s) \right| &\leqq \left(\sum_{i=1}^{\infty} (\varphi_i, v)^2 \right)^{\frac{1}{2}} \operatorname*{Max}_{s, s+h \in \bar{\Omega}} \left(\sum_{i=1}^{\infty} \frac{(\varphi_i(s+h) - \varphi_i(s))^2}{\lambda_i} \right)^{\frac{1}{2}} \\ &\leqq \frac{2\sqrt{2}}{\pi} \left(\sum_{i=1}^{\infty} \frac{\sin^2 i \frac{\pi}{2} h}{i^2} \right)^{\frac{1}{2}} \|v\|_{L^2[\Omega]}. \end{aligned}$$

Die Reihe $\sum_{i=1}^{\infty} s_i(h)$ mit $s_i(h) = \frac{1}{i^2} \sin^2 i \frac{\pi}{2} h$ ist absolut und gleichmäßig konvergent, und für jeden einzelnen Summanden gilt $s_i(h) \to 0$ $(i = 1, 2, 3 \ldots)$ für $h \to 0$. Durch Vertauschung der Reihenfolge der Grenzwerte folgt daher

$$\lim_{h \to 0} \operatorname*{Max}_{s, s+h \in \bar{\Omega}} \left| (K_1^{\frac{1}{2}} v)(s+h) - (K_1^{\frac{1}{2}} v)(s) \right| \leqq \frac{2\sqrt{2}}{\pi} \left(\sum_{i=1}^{\infty} \lim_{h \to 0} s_i(h) \right)^{\frac{1}{2}} \|v\|_{L^2[\Omega]} = 0.$$

Es gilt daher $K_1^{\frac{1}{2}} v \in C^0[\bar{\Omega}] = B_1$. Mittels der Schwarzschen Ungleichung erhält man ferner

$$\begin{aligned} \|K_1^{\frac{1}{2}} v\|_{C^0[\bar{\Omega}]} = \operatorname*{Max}_{s \in \bar{\Omega}} \left| \sum_{i=1}^{\infty} \frac{(\varphi_i, v)}{\sqrt{\lambda_i}}\, \varphi_i(s) \right| &\leqq \left(\sum_{i=1}^{\infty} (\varphi_i, v)^2 \right)^{\frac{1}{2}} \operatorname*{Max}_{s \in \bar{\Omega}} \left(\sum_{i=1}^{\infty} \frac{(\varphi_i(s))^2}{\lambda_i} \right)^{\frac{1}{2}} \\ &\leqq \frac{\sqrt{2}}{\pi} \left(\sum_{i=1}^{\infty} \frac{1}{i^2} \right)^{\frac{1}{2}} \|v\|_{L^2[\Omega]}, \end{aligned}$$

womit (III′) vollständig bewiesen ist.

(IV′) trivial.

(V′) Eine elementare Rechnung zeigt, daß $\tilde{L} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$ die Bedingungen von Voraussetzung (V′) erfüllt.

(VI′) Sei $u_j \in C^0[\bar{\Omega}]$ $(j = 1, 2)$, dann folgt wegen (4.17)

$$\begin{aligned}\big(J_2 M_1(u_1) - J_2 M_1(u_2),\, J_1 u_1 - J_1 u_2\big) &= \int_{\Omega} \big(g(s, u_1(s)) - g(s, u_2(s))\big) \\ &\cdot \big(u_1(s) - u_2(s)\big)\, ds \leqq a \int_{\Omega} \big(u_1(s) - u_2(s)\big)^2 ds \leqq a \,\|J_1(u_1 - u_2)\|^2_{L^2[\Omega]},\end{aligned}$$

womit (3.11) bewiesen ist. (3.12) folgt aus (4.18) unter Beachtung von

$$\|K^{\frac{1}{2}}\|_{\mathcal{L}^2 \to \mathcal{L}^2} = \frac{1}{\sqrt{\lambda_1}} = \frac{1}{\pi}.$$

(VII′) Sei $u \in C^0[\bar{\Omega}]$, dann gilt nach (4.20)

$$\|M_2(u)\|_B = \operatorname{Max}\big(|F_1(u)|, |F_2(u)|\big) \leqq b.$$

(VIII′) $\tilde{L} M_2$ ist kompakt von $C^0[\bar{\Omega}]$ in $B = R^2$, da R^2 endlich dimensional ist. Die Voraussetzungen von Satz 2 sind erfüllt. Es existiert daher mindestens eine „verallgemeinerte" Lösung von (4.16). Damit ist Lemma 5 bewiesen.

Nach Definition 1 und Lemma 5 existieren daher ein $y^* = \{y_1^*, y_2^*\} \in R^2$ und ein $u^* \in C^0[\bar{\Omega}]$ mit

$$u^*(s) = y_1^* + y_2^* s + \int_0^1 K(s, t)\, g\big(t, u^*(t)\big)\, dt, \tag{4.23a}$$

$$y_1^* = F_1(u^*), \qquad y_1^* + y_2^* = F_2(u^*). \tag{4.23b}$$

Wegen $u^* \in C^0[\bar{\Omega}]$ und Voraussetzung (a′) gilt $g\big(s, u^*(s)\big) \in C^0[\Omega] \cap L^2[\Omega] = \tilde{B}_0$. Aus (II′) folgt hieraus $K M_1(u^*) \in D(L_1)$. Zusammen mit Bemerkung 2 ist $u^*(s)$ eine „klassische" Lösung von (4.16).

Zusammenfassend erhalten wir den

Satz 5. Es seien die Voraussetzungen (a′) und (b′) erfüllt. Dann gibt es mindestens eine Funktion $u^*(s) \in C^0[\bar{\Omega}] \cap C^2[\Omega]$, die das Problem (4.16) (im klassischen Sinn) löst.

Bemerkung 7a. Es seien $\alpha_1, \alpha_2, \ldots, \alpha_{11}$ reelle Zahlen, dann erfüllen beispielsweise die Abbildungen

$$F_1(u) = \alpha_1 + \frac{\alpha_2}{1 + \alpha_3^2 (u(0))^2 + \alpha_4^2 \exp\big(u(\tfrac{1}{2})\big) + \alpha_5^2 \int_0^1 (u(s))^2\, ds},$$

$$F_2(u) = \alpha_6 + \frac{\alpha_7 (u(0))^3 + \alpha_8}{1 + \alpha_9^2 (u(0))^4 + \alpha_{10}^2 (u(1))^2 + \alpha_{11}^2 \operatorname{Max}_{0 \leqq s \leqq 1} |u(s)|}$$

die Voraussetzung (b′).

Bemerkung 7b. Es seien $a(s)$, $b(s)$ und $c(s)$ auf $[0, 1]$ stetige Funktionen mit $b(s) \geqq 0$ und $c(s) \geqq 0$. Ferner seien die Konstanten $\beta_i < \frac{1}{2}$ $(i = 1, 2, 3, 4)$, und es sei

$$g(s, u) = \frac{a(s)}{s^{\beta_1}} - \frac{b(s)}{(1-s)^{\beta_2}} \sinh u - \frac{c(s)}{s^{\beta_3}(1-s)^{\beta_4}} \frac{u^{2m-1}}{1+u^2}$$

mit $m \geqq 2$ und ganzzahlig. Dann ist die Voraussetzung (a') erfüllt.

Bemerkung 8. Die Probleme unter Bemerkung 6 (mit $d(s) \not\equiv 0$ und $m \geqq 3$) und Bemerkung 7 (mit $b(s) \not\equiv 0$ oder $c(s) \not\equiv 0$ und $m \geqq 3$) können mit Hilfe der Existenzsätze von Ehrmann [3], [4] und Conti [1] nicht behandelt werden.

Literatur

1. Conti, R.: Problemi quasi lineari negli spazi di Banach. Rend. Accad. Naz. Lincei **32**, Ser. 8, 495—498 (1962).
2. Courant, R., Hilbert, D.: Methoden der Mathematischen Physik. Berlin: Springer Bd. I 1931 und Bd. II 1937.
3. Ehrmann, H.: Ein Existenzsatz für die Lösungen gewisser Gleichungen mit Nebenbedingungen bei beschränkter Nichtlinearität. Arch. Rational Mech. Anal. 7, 349—358 (1961).
4. — Existenzsätze für die Lösungen gewisser nichtlinearer Randwertaufgaben. ZAMM **45**, T 22—T 29 (1965).
5. Günter, N. M.: Die Potentialtheorie und ihre Anwendung auf Grundaufgaben der Mathematischen Physik. Leipzig: B. G. Teubner 1957.
6. Horn, J.: Partielle Differentialgleichungen. Berlin: Walter de Gruyter & Co. 1949.
7. Krasnoselskii, M. A.: Topological methods in the theory of nonlinear integral equations. Oxford-London-New York-Paris: Pergamon Press 1964.
8. Petry, W.: Nichtlineare Operator- und Integralgleichungssysteme. Erscheint in Math. Nachr.
9. Riesz, F., Sz.-Nagy, B.: Vorlesungen über Funktionalanalysis. Berlin: Deutscher Verlag der Wissenschaften 1956.
10. Vainberg, M. M.: Variational methods for the study of nonlinear operators. San Francisco-London-Amsterdam: Holden-Day, Inc. 1964.

W. Petry
Kernforschungsanlage Jülich
Zentralinstitut für Angewandte Mathematik
5170 Jülich, Postfach 365

Numer. Math. 14, 495—507 (1970)

On the Bauer's Scaled Condition Number of Matrices Arising from Approximate Conformal Mapping

Hana Švecová

Received July 10, 1969

Abstract. Approximation of a conformal mapping of a simply connected domain onto a circle by means of Ritz's method yields a system of linear equations with a Gram matrix. The asymptotic behaviour of the minimal condition of this matrix is studied in dependence on its order.

I.

Introduction. The problem to determine numerically a conformal mapping of a bounded simply connected domain G onto a circle can, with the aid of extremal functions, be reduced to the solution of a sequence of finite systems of linear algebraic equations with Gram matrices A_n $(n=1, 2, \ldots)$, the order of A_n being n. The elements of these matrices depend upon the definition of the Hilbert space H where the solution is sought and upon the basis chosen in H.

The influence of rounding errors on the numerical solution of the system $A_n x = b$ (especially for Gauss elimination in floating-point computation) may be studied in terms of Bauer's scaled condition number

$$\mathscr{B}(A_n) = \inf_{D_1, D_2} \operatorname{cond}(D_1 A_n D_2) \tag{1}$$

where $\operatorname{cond}(B) = \|B\| \, \|B^{-1}\|$ and D_1, D_2 are diagonal matrices of order n with positive diagonal elements (for real matrices see [1, 4]). By $\|B\|$ we shall denote the spectral norm of B. In order to maintain the Hermitian symmetry of the matrix, we shall moreover suppose $D_1 = D_2$. This is in fact no restriction, since the value of the infimum in (1) does not change (for proof see [3]).

The aim of this paper is to contribute to the solution of the problem if there exists a reasonably computable basis of a given Hilbert space of analytic functions satisfying the inequality

$$\sup \mathscr{B}(A_n) < \infty.$$

A positive answer to this question is obvious if G is a disk and $H = L_2(G)$. However, as we shall show, this inequality does not hold for domains G with boundaries arbitrarily close to the circle, e.g. ellipses and polygons, with the same definition of base functions.

The Variational Method. Let G be a bounded simply connected domain in the complex plane, $0 \in G$. Let $w = f(z)$ be the conformal mapping of G onto a disk with center at the coordinate origin, satisfying the conditions $f(0) = 0$, $f'(0) = 1$. (Then the radius R of the disk is uniquely determined.)

Let $L_2(G)$ denote the set of all the functions F which are regular in G and satisfy (using notation $z = x + iy$)

$$\|F\|^2 = \iint_G |F(z)|^2 \, dx \, dy < \infty .$$

If $F_1, F_2 \in L_2(G)$, we can define the scalar product

$$(F_1, F_2) = \iint_G F_1(z) \, \overline{F_2(z)} \, dx \, dy .$$

By this definition the set $L_2(G)$ becomes a Hilbert space (see [5], p. 117).

It follows from the least area principle of the theory of functions (see [2], p. 189, [5], p. 119, [6], p. 281) that the derivative $f'(z)$ is the only solution to the problem of minimization of the functional

$$\iint_G |F(z)|^2 \, dx \, dy$$

on the set of functions $F \in L_2(G)$ satisfying $F(0) = 1$.

The derivative $f'(z)$ can be approximated by means of the Ritz's method ([5], p. 120; [6], p. 383). We shall give here a slightly more general formulation of this method than the mentioned monographs do. However, it is easy to see that the generalization does not influence the validity of the assertions cited without proof.

Let $\{u_{k,n}(z)\}$ $(k = 0, 1, \ldots, n;\ n = 1, 2, \ldots)$ be a complete system of functions in $L_2(G)$, such that for each n the functions $u_{k,n}(z)$ $(k = 0, 1, \ldots, n)$ are linearly independent. Moreover, let $u_0(0) \neq 0$. Then the n-th approximation of the derivative $f'(z)$ is defined to be the function $\varphi_n(z)$ which minimizes the functional $\|F\|$ in the class of all complex linear combinations of the functions $u_0, u_1, \ldots, u_n$ satisfying $F(0) = 1$. The function φ_n is uniquely determined by the relation

$$(\varphi_n, \psi_n) = 0$$

which is satisfied by any linear combination ψ_n of the functions $u_{k,n}$ fulfilling the condition $\psi_n(0) = 0$. If we write

$$\varphi_n = \sum_{k=0}^{n} a_{k,n} u_{k,n} ,$$

then, under additional assumptions

$$u_{0,n}(0) = 1 , \qquad u_{k,n}(0) = 0 \quad (k = 1, 2, \ldots, n) ,$$

the coefficients $a_{k,n}$ $(k = 1, 2, \ldots, n)$ are given by the solution of a system of n linear equations

$$\sum_{k=1}^{n} (u_{k,n}, u_{j,n}) \, a_{k,n} = -(u_{0,n}, u_{j,n}) \quad (j = 1, 2, \ldots, n) \tag{2}$$

with a Gram matrix $A_n = ((u_{k,n}, u_{j,n}))$. Hence A_n is a positive definite Hermitian matrix for each n.

In practical computations, it is common to use the polynomial basis of $L_2(G)$, i.e. to take $u_k(z) = z^k$ $(k = 0, 1, \ldots)$. (We have dropped the subscript n, since

the base functions do not depend on n.) If G is a disk with center at the origin and radius R, then the matrix A_n of system (2), generated by the polynomial basis, is a diagonal matrix with diagonal elements $R^{2(k+1)}/(k+1)$ $(k=1, 2, \ldots, n)$, hence (denoting the maximal and minimal eigenvalue of A_n by $\lambda_{\max}(A_n)$ and $\lambda_{\min}(A_n)$, respectively)

$$\lim_{n\to\infty} \operatorname{cond}(A_n) = \lim_{n\to\infty} \frac{\lambda_{\max}(A_n)}{\lambda_{\min}(A_n)} = \infty.$$

However, if we prefer the base functions $v_k(z) = C_k z^k$ with $C_k = \sqrt{(k+1)/R^k}$ $(k=0, 1, \ldots)$, then, denoting $B_n = ((v_k, v_j))$, we have

$$\operatorname{cond}(B_n) = 1$$

for each n.

Obviously, the replacement of the system $\{u_k\}$ by $\{v_k\}$ is equivalent to the scaling of A_n by the diagonal matrix $D_n(C_1, \ldots, C_n)$, i.e. to the replacement of A_n by $D_n A_n D_n$.

Hence if G is a disk with center at the origin and the matrices A_n are generated by the polynomial basis, we have

$$\sup_n \mathscr{B}(A_n) < \infty.$$

In the following we shall show that this inequality does not hold for a large class of domains not only for the polynomial basis, but for a more general class of bases $u_0 = 1$, $u_{k,n} = C_{k,n} z^k$ $(k=1, 2, \ldots, n;\ n=1, 2, \ldots)$, where $C_{k,n}$ are arbitrary complex numbers.

II.

For further need, we recall the following

Lemma. Let A be a positive definite Hermitian matrix and B its principal submatrix. Then we have

$$\lambda_{\min}(A) \leqq \lambda_{\min}(B),$$
$$\lambda_{\max}(A) \geqq \lambda_{\max}(B).$$

Definition. A domain G is said to be starlike with respect to a point ζ if each ray initiating at ζ intersects the boundary of G exactly at one point.

Theorem. Let G be a starlike domain with respect to the coordinate origin, situated inside the unit circle and such that its boundary touches the unit circle exactly at a finite number of points. Let the functions $u_{k,n}$ have the form $C_{k,n} z^k$ with arbitrary non-zero complex constants $C_{k,n}$ $(k=0, 1, \ldots, n;\ n=1, 2, \ldots)$. Let A_n be the matrix of the system (2). Then

$$\sup_n \mathscr{B}(A_n) = \infty$$

where $\mathscr{B}(A_n)$ is the Bauer's scaled condition number (1) of the matrix A_n.

Proof. a) Notice first that the system $u_{k,n}$ is complete in consequence of the starlikeness of the domain G (see [7], p. 428).

Suppose that the theorem does not hold, i.e. that

$$\sup_n \mathscr{B}(A_n) < \infty. \tag{3}$$

The value of $\mathscr{B}(A_n)$ obviously does not decrease with increasing n, hence

$$\sup_n \mathscr{B}(A_n) = \lim_{n\to\infty} \mathscr{B}(A_n).$$

Since A_n is a positive definite Hermitian matrix, we have

$$\operatorname{cond}(A_n) = \|A_n\| \, \|A_n^{-1}\| = \lambda_{\max}(A_n)/\lambda_{\min}(A_n).$$

The same holds for the scaled matrices $D_n A_n D_n$.

We can choose matrices D_n^* $(n = 1, 2, \ldots)$ so that

$$\lim_{n\to\infty} \mathscr{B}(A_n) = \lim_{n\to\infty} \frac{\lambda_{\max}(D_n^* A_n D_n^*)}{\lambda_{\min}(D_n^* A_n D_n^*)}.$$

Denote by $d_{k,n}$ the k-th element of the principal diagonal of the matrix D_n^*. Let $\alpha_{jk}^{(n)}$ $(j, k = 1, \ldots, n)$ denote the elements of the matrix $D_n^* A_n D_n^*$, i.e.

$$\alpha_{jk}^{(n)} = d_{j,n} d_{k,n} \overline{C_{j,n}} C_{k,n}(z^k, z^j).$$

We can write

$$\alpha_{jk}^{(n)} = \gamma_{j,n} \gamma_{k,n} \beta_{j,k} \tag{4}$$

where

$$\beta_{jk} = \frac{(z^k, z^j)}{\|z^k\| \, \|z^j\|} \tag{5}$$

is independent of n and

$$\gamma_{k,n} = d_{k,n} C_{k,n} \|z^k\|.$$

The assumption (3) implies the existence of two positive numbers m, M such that we have

$$m \leqq \lambda_{\min}(D_n^* A_n D_n^*) \leqq \lambda_{\max}(D_n^* A_n D_n^*) \leqq M \tag{6}$$

for all n. The diagonal elements of the matrix $D_n^* A_n D_n^*$ have the form

$$\alpha_{kk}^{(n)} = |\gamma_{k,n}|^2 \quad (k = 1, \ldots, n).$$

Hence, according to the preceding Lemma and to the relation (4), we have

$$m \leqq |\gamma_{k,n}|^2 \leqq M \tag{7}$$

for $k = 1, \ldots, n$ and $n = 1, 2, \ldots$.

Let B denote the matrix of infinite order with elements β_{jk} $(j, k = 1, 2, \ldots)$ defined by (5). We shall prove the existence of a sequence $\{\beta_{j_\nu k_\nu}\}$ of non-diagonal elements of the matrix B, such that

$$\lim_{\nu\to\infty} |\beta_{j_\nu k_\nu}| = 1. \tag{8}$$

Since all the diagonal elements of the matrix B are equal to 1, this will enable us to construct a sequence $\{X_\nu\}$ of second-order principal submatrices of B con-

verging to a matrix

$$\begin{pmatrix} 1 & \delta \\ \bar{\delta} & 1 \end{pmatrix}$$

satisfying $|\delta| = 1$. To the sequence $\{X_\nu\}$, we can choose a sequence $\{Y_\nu\}$ of equally situated second-order principal submatrices of matrices $D^*_{p_\nu} A_{p_\nu} D^*_{p_\nu}$, where the sequence $\{p_\nu\}$ satisfies the conditions

$$p_\nu \geqq \max(i_\nu, k_\nu),$$

$$\lim_{\nu \to \infty} p_\nu = \infty.$$

Then the relations (4), (7), (8) imply the existence of a subsequence of $\{Y_\nu\}$ converging to a matrix

$$\begin{pmatrix} |\delta_1|^2 & \bar{\delta}_1 \delta_2 \delta \\ \delta_1 \bar{\delta}_2 \bar{\delta} & |\delta_2|^2 \end{pmatrix}$$

where δ_1, δ_2 are limits of certain subsequences of the sequences $\{\gamma_{j_\nu, p_\nu}\}$ and $\{\gamma_{k_\nu, p_\nu}\}$, respectively. Obviously, the minimal eigenvalues of submatrices belonging to the mentioned subsequence converge to the minimal eigenvalue of the limiting matrix, i.e. to zero. This implies (by the Lemma) the relation

$$\inf_n \lambda_{\min}(D^*_n A_n D^*_n) = 0$$

which contradicts the assumption (3).

b) Using polar coordinates to describe the boundary of the domain G, we can write $r = r(\varphi)$ (where $\varphi = \arg z$, $r = |z|$). Then we have

$$\beta_{jk} = \frac{(z^k, z^j)}{\|z^k\| \, \|z^j\|} = \sqrt{b_{jk}} Q_{jk} \tag{9}$$

where

$$b_{jk} = \frac{4(j+1)(k+1)}{(j+k+2)^2},$$

$$Q_{jk} = \frac{\int_0^{2\pi} [r(\varphi)]^{j+k+2} \cos[(k-j)\varphi] \, d\varphi + i \int_0^{2\pi} [r(\varphi)]^{j+k+2} \sin[(k-j)\varphi] \, d\varphi}{\left\{ \int_0^{2\pi} [r(\varphi)]^{2(j+1)} d\varphi \cdot \int_0^{2\pi} [r(\varphi)]^{2(k+1)} \, d\varphi \right\}^{\frac{1}{2}}}.$$

Denote the difference $k-j$ by L. In part b of the proof, we will consider L as a fixed number and we will find the limit

$$\lim_{j \to \infty} \operatorname{Re} \beta_{j, j+L}.$$

We shall use the following notation.

$$P_+(x) = \frac{\int_0^{2\pi} [r(\varphi)]^L \cos(L\varphi) \cdot [r(\varphi)]^x \, d\varphi}{\int_0^{2\pi} [r(\varphi)]^x \, d\varphi},$$

$$P_-(x) = \frac{\int_0^{2\pi} [r(\varphi)]^{-L} \cos(L\varphi) \cdot [r(\varphi)]^x \, d\varphi}{\int_0^{2\pi} [r(\varphi)]^x \, d\varphi}.$$

It is easy to see that

$$(\mathrm{Re}\,\beta_{j,j+L})^2 = b_{j,j+L} P_+\big(2(j+1)\big) P_-\big(2(j+L+1)\big). \tag{10}$$

For $j \to \infty$, we have

$$b_{j,j+L} = 1 + O(1/j^2).$$

We shall now study the rest of the expression in (10).

Let m denote the number of common boundary points of the domain G and the unit circle and let ψ_ν $(\nu = 1, 2, \ldots, m)$ denote the arguments of these points. Thus we have $r(\psi_\nu) = 1$ $(\nu = 1, 2, \ldots, m)$ and $r(\varphi) < 1$ if

$$\varphi \notin \bigcup_{\nu=1}^{m} \psi_\nu.$$

First, σ being an arbitrary positive number, we shall construct a system of intervals $J_\nu(\sigma, x)$ $(\nu = 1, 2, \ldots, m;\ x > 0)$ satisfying the following three conditions.

(i) $\psi_\nu \in J_\nu(\sigma, x)$;

(ii) if $d[J]$ denotes the length of the interval J, then for any $x > 0$ we have

$$\frac{\sigma}{m} \leqq \sum_{\nu=1}^{m} d[J_\nu(\sigma, x)] \leqq \sigma;$$

(iii) $$\int_{J_\nu(\sigma, x)} r^x \, d\varphi = \int_{J_\mu(\sigma, x)} r^x \, d\varphi$$

for $\mu, \nu = 1, 2, \ldots, m$ and all $x > 0$.

Define $g_\nu(y)$ $(\nu = 1, \ldots, m)$ by

$$g_\nu(y) = r(\psi_\nu + y).$$

For a moment, consider x fixed. Let μ satisfy the following equality:

$$\min_\nu \int_0^{\frac{\sigma}{2m}} [g_\nu(y)]^x \, dy = \int_0^{\frac{\sigma}{2m}} [g_\mu(y)]^x \, dy = a(\sigma, x).$$

The continuous dependence of the integral on its upper bound of integration implies the existence of numbers $\sigma_\nu \leqq \sigma/(2m)$ such that for $\nu \neq \mu$ we have

$$\int_0^{\sigma_\nu} [g_\nu(y)]^x \, dy = a(\sigma, x).$$

Moreover, define σ_μ by $\sigma_\mu = \sigma/2m$. Now we can construct the right-hand sections of the sought intervals:

$$J_{\nu_+}(\sigma, x) = \langle \psi_\nu, \psi_\nu + \sigma_\nu \rangle$$

$(\nu = 1, \ldots, m)$. The left-hand sections $J_{\nu_-}(\sigma, x)$ can be constructed in a similar way, and finally we take

$$J_\nu(\sigma, x) = J_{\nu_+}(\sigma, x) \cup J_{\nu_-}(\sigma, x)$$

$(\nu = 1, \ldots, m)$. It is easily seen that these intervals satisfy the conditions (i) to (iii).

In the following, we shall assume that σ is sufficiently small to make the intervals mutually disjoint.

Let

$$a(\sigma, x) = \int_{J_\nu} r^x \, d\varphi \quad (\nu = 1, \ldots, m).$$

We shall use the following notation:

$$J = J(\sigma, x) = \bigcup_{\nu=1}^{m} J_\nu(\sigma, x),$$

$$K = K(\sigma, x) = \langle 0, 2\pi \rangle - J(\sigma, x).$$

Then we can write

$$P_\pm(x) = \frac{\sum_{\nu=1}^{m} \frac{\int_{J_\nu} g_\pm(\varphi)\, r^x \, d\varphi}{m\, a(\sigma, x)} + \frac{\int_K g_\pm(\varphi)\, r^x \, d\varphi}{\int_J r^x \, d\varphi}}{1 + \frac{\int_K r^x \, d\varphi}{\int_J r^x \, d\varphi}}, \tag{11}$$

where $g_\pm(\varphi) = r(\varphi)^{\pm L} \cos L\varphi$. By the sign $\pm$ we mean that one of the signs $+$ and $-$ is valid on both sides of the equality. We shall also write g instead of $g_\pm$.

By the first mean-value theorem, there exist numbers $\varphi_\nu(\sigma, x)$ $(\nu = 1, \ldots, m)$ (generally different for P_+ and P_-) such that

$$\int_{J_\nu} g(\varphi)\, r^x \, d\varphi = g[\varphi_\nu(\sigma, x)] \int_{J_\nu} r^x \, d\varphi = g[\varphi_\nu(\sigma, x)]\, a(\sigma, x).$$

Hence the first term of the numerator in (11) is equal to

$$\frac{1}{m} \sum_{\nu=1}^{m} g[\varphi_\nu(\sigma, x)].$$

Let

$$M = \max_{\varphi \in \bar{K}} r(\varphi).$$

We have $M < 1$, hence for each ν $(\nu = 1, \ldots, m)$ there exists a set $J_\nu^* \subset J_\nu$ of positive measure so that

$$\inf_{\varphi \in J_\nu^*} r(\varphi) > M.$$

Define J^* by

$$J^* = \bigcup_{\nu=1}^{m} J_\nu^*$$

and let $\mu(J^*)$ be the Lebesgue measure of the set J^*. If we write

$$N = \inf_{\varphi \in J^*} r(\varphi),$$

$$Q = \max_{\varphi \in \bar{K}} g(\varphi)$$

(Q is independent of x), we are led to the conclusion

$$\left|\frac{\int_K g(\varphi)\, r^x\, d\varphi}{\int_J r^x\, d\varphi}\right| \leqq \frac{Q(2\pi-\sigma/m)\, M^x}{\mu(J^*)\, N^x} \xrightarrow[x\to\infty]{} 0. \tag{12}$$

Similarly we have

$$\frac{\int_K r^x\, d\varphi}{\int_J r^x\, d\varphi} \leqq C(\sigma)\,(M/N)^x \xrightarrow[x\to\infty]{} 0. \tag{13}$$

In the following, $P(x)$ will denote either $P_+(x)$ or $P_-(x)$. Let us now take σ as a fixed number. For each ν, all the numbers $\varphi_\nu(\sigma, x)$ belong to a closed interval the length of which does not exceed σ. Hence we can find a sequence $\{x_n\}$ and numbers $\omega_\nu(\sigma)$ such that

$$\lim_{n\to\infty} P(x_n) = \limsup_{x\to\infty} P(x)$$

and

$$\lim_{n\to\infty} \varphi_\nu(\sigma, x_n) = \omega_\nu(\sigma)$$

for $\nu = 1, 2, \ldots, m$. Then by (11), (12), (13) we have

$$\limsup_{x\to\infty} P(x) = \lim_{n\to\infty} P(x_n) = \lim_{n\to\infty} \left\{1/m \sum_{\nu=1}^{m} g\,[\varphi_\nu(\sigma, x_n)]\right\} = 1/m \sum_{\nu=1}^{m} g\,[\omega_\nu(\sigma)].$$

However, $P(x)$ does not depend on σ, so that we have

$$\limsup_{x\to\infty} P(x) = 1/m \sum_{\nu=1}^{m} g\left[\lim_{\sigma\to\infty} \omega_\nu(\sigma)\right] = 1/m \sum_{\nu=1}^{m} g(\psi_\nu).$$

In a similar way, we can prove

$$\liminf_{x\to\infty} P(x) = 1/m \sum_{\nu=1}^{m} g(\psi_\nu),$$

whence (since $r(\psi_\nu) = 1$) it follows that

$$\lim_{x\to\infty} P_\pm(x) = 1/m \sum_{\nu=1}^{m} g_\pm(\psi_\nu) = 1/m \sum_{\nu=1}^{m} \cos L\psi_\nu.$$

Returning to (10), we obtain

$$\lim_{j\to\infty} |\operatorname{Re} \beta_{j,j+L}| = 1/m \left|\sum_{\nu=1}^{m} \cos L\psi_\nu\right|. \tag{14}$$

Remark. This already proves the theorem under the additional assumption that ψ_ν/π is rational for $\nu = 1, \ldots, m$. For example, if G is the interior of an ellipse with $\psi_1 = 0$ and $\psi_2 = \pi$, we have (setting $L = 2$)

$$\lim_{j\to\infty} |\operatorname{Re} \beta_{j,j+2}| = 1.$$

Since $|\beta_{jk}| \leqq 1$ for any j, k, we thus obtain relation (8) which (by Section a) of this proof) proves the theorem.

c) Now it is left to prove the existence of a sequence $\{L_n\}$ of natural numbers, such that

$$\lim_{n\to\infty} \cos(L_n \psi_\nu) = 1 \tag{15}$$

for $\nu = 1, \ldots, m$. This will enable us, by combining (14) and (15), to find sequences $\{j_n\}$ and $\{k_n\}$ satisfying

$$\lim_{n\to\infty} |\operatorname{Re} \beta_{j_n k_n}| = 1$$

and hence also

$$\lim_{n\to\infty} |\beta_{j_n k_n}| = 1 .$$

Let us first assume $m = 1$ and write simply ψ instead of ψ_1. We shall show that there exists a sequence $\{L_n, K_n\}$ of pairs of natural numbers, such that

$$\lim_{n\to\infty} |L_n \psi - 2K_n \pi| = 0. \tag{16}$$

Writing $\omega = 2\pi/\psi$, we have

$$|L\psi - 2K\pi| = |\psi| \cdot |L - K\omega|$$

for any L, K. We can construct the sought sequence by the following iterative process.

Let L_0, K_0 be natural numbers such that

$$K_0 = L_0 + \delta_0,$$

where $0 < |\delta_0| < 1$. Let us assume $\delta_0 > 0$. For a negative δ_0 we can proceed quite analogously. Then for any j we have

$$j K_0 \omega = j L_0 + j \delta_0 .$$

Let us define j_0 by $j_0 = [\delta_0^{-1}]$. Hence we have

$$j_0 \delta_0 \leqq 1, \qquad (j_0 + 1)\, \delta_0 > 1 .$$

If $j_0 \delta_0 = 1$, then we can define $K_n = j_0 K_0$, $L_n = j_0 L_0 + 1$ $(n = 1, 2, \ldots)$ to satisfy (16). Therefore we shall now assume $j_0 \delta_0 < 1$. We have

$$j_0 K_0 \omega = j_0 L_0 + 1 - d_1,$$
$$(j_0 + 1) K_0 \omega = (j_0 + 1) L_0 + 1 + d_2$$

where

$$d_1 = 1 - j_0 \delta_0 > 0, \qquad d_2 = (j_0 + 1)\, \delta_0 - 1 > 0.$$

Obviously $d_1 + d_2 = \delta_0$, whence $\min(d_1, d_2) \leqq \delta_0/2$. Now if $d_1 < d_2$, we set

$$K_1 = j_0 K_0,$$
$$L_1 = j_0 L_0 + 1,$$
$$\delta_1 = -d_1.$$

In the other case, i.e. if $d_1 \geqq d_2$, we set

$$K_1 = (j_0 + 1) K_0,$$
$$L_1 = (j_0 + 1) L_0 + 1,$$
$$\delta_1 = d_1.$$

Thus we come to the relation

$$K_1 \omega = L_1 + \delta_1, \quad |\delta_1| \leqq \delta_0/2.$$

If we started with a negative δ_0, we should obtain the same relation.

In the following step, K_1, L_1 and δ_1 take the place of K_0, L_0 and δ_0, respectively. By this process, we obtain a sequence $\{K_n, L_n\}$ such that

$$\lim_{n\to\infty} |L_n \psi - K_n \pi| \leqq \lim_{n\to\infty} (|\psi| \cdot \delta_0/2^n) = 0$$

which is equivalent to (15) for $\nu = 1$.

Now let $m > 1$. Let us assume that there exists a sequence $\{L_n\}$ satisfying

$$\lim_{n\to\infty} \cos L_n \psi_\nu = 1 \tag{17}$$

for $\nu = 1, \ldots, m-1$. We claim that (17) is then true for $\nu = 1, \ldots, m$.

We shall use the following notation. $\tilde{\varphi} = \varphi - 2k\pi$, k being an integer such that $\tilde{\varphi} \in (-\pi, \pi)$, and $\hat{\varphi} = \varphi \pmod{2\pi}$.

Let us consider the following two statements.

(i) There exists a sequence $\{L_n\}$ of natural numbers, such that

$$\lim_{n\to\infty} \max_{\nu=1,\ldots,m-1} |\widetilde{L_n \psi_\nu}| = 0. \tag{18}$$

(ii) To each sequence $\{L_n\}$ enjoying the property (18), there exists a number $a > 0$, such that

$$|\widetilde{L_n \psi_m}| > a \tag{19}$$

for all n.

If we prove that these two statements contradict each other, it will imply that if condition (i) is true, then it holds also if we replace (18) by

$$\lim_{n\to\infty} \max_{\nu=1,\ldots,m} |\widetilde{L_n \psi_\nu}| = 0.$$

It is easy to see that this is equivalent to (17) for $\nu = 1, \ldots, m$.

To this end, considering the conditions (i) and (ii) as being true, let $\{L_n\}$ be a sequence satisfying (18) and let $a > 0$ be the number corresponding to this sequence by condition (ii).

As a consequence of (i), to any natural number k there exists a number $N(k)$, so that we have

$$\max_{\nu=1,\ldots,m-1} |\widetilde{L_n \psi_\nu}| \leqq a/k^2$$

for $n \geqq N(k)$. We can assume that

$$\lim_{n\to\infty} N(k) = \infty.$$

Then for $\mu = 1, 2, \ldots, k$ we have

$$\max_{\nu=1,\ldots,m-1} |\widetilde{\mu L_{N(k)} \psi_\nu}| \leqq a/k.$$

Let us construct a sequence

$$\{t_n\} = L_{N(1)}, L_{N(2)}, 2L_{N(2)}, L_{N(3)}, 2L_{N(3)}, 3L_{N(3)}, \ldots,$$

where for $n = k(k-1)/2 + \mu$, $\mu = 1, 2, \ldots, k$ we have

$$t_n = \mu L_{N(k)}.$$

Hence to any natural number k there exists a number $N(k)$ such that

$$\max_{\nu=1,\ldots,m-1} |\widetilde{t_n \psi_\nu}| \leqq a/k$$

for $n > N^*(k)$. On the other side, by (ii), there exists a positive number $b \leqq a$ so that

$$|t_n \psi_m| > b$$

for all n.

Let c be a natural number such that $a/c \leqq b$. We shall select those elements of the sequence $N(k)$ where $k = qc$ with a natural q and form the following subsequence of $\{t_n\}$:

$$\{v_n\} = L_{N(c)}, L_{N(2c)}, 2L_{N(2c)}, L_{N(3c)}, 2L_{N(3c)}, 3L_{N(3c)}, \ldots,$$

i.e. for $n = q(q-1)/2 + \mu$, $\mu = 1, 2, \ldots, q$, we have $v_n = \mu L_{N(qc)}$.

Obviously, we have

$$\max_{\nu=1,\ldots,m-1} |\widetilde{v_n \psi_\nu}| \leqq a/k = a/(qc) \leqq b/q$$

starting with a certain n_q for any natural q. At the same time, we have

$$|\widetilde{v_n \psi_m}| > b \tag{20}$$

for all n.

If the inequality (20) holds, then, having in mind the structure of the sequence $\{v_n\}$, we realize that the following assertion must be valid:

(iii) To any natural number q there exists an integer ω_q such that the inequality

$$|\widetilde{\mu \omega_q \psi_m}| > b$$

holds for $\mu = 1, 2, \ldots, q$.

We will show that no ψ_m enjoys the property (iii).

Let φ be an arbitrary number belonging to the interval $(b, 2\pi - b)$. Let $\varkappa$, N be defined by

$$\varkappa = [2\pi/b] + 1, \qquad N = [(\log \pi - \log b)/\log 2], \tag{21}$$

square brackets denoting the whole part of the number. Let us assume that

$$|\widetilde{j\varphi}| > b \tag{22}$$

for $j = 1, 2, \ldots, q_0$ where $q_0 \geqq \varkappa^N$.

Let $k_0^* = [2\pi/\varphi]$. Hence we have $k_0^* < \varkappa - 1$. By assumption (22) we have

$$k_0^* < 2\pi - b, \quad \widehat{(k_0^* + 1)\varphi} > b.$$

Further let $\delta_0 = \min(d_1, d_2)$ where

$$d_1 = 2\pi - k_0^* \varphi, \quad d_2 = (k_0^* + 1)\varphi.$$

We have $d_1 + d_2 = \varphi$, whence

$$\delta_0 \leqq \varphi/2 < (2\pi - b)/2.$$

Let us define k_0 by

$$k_0 = \begin{cases} k_0^* & \text{if} \quad d_1 < d_2 \\ k_0^* + 1 & \text{if} \quad d_1 \geqq d_2. \end{cases}$$

Obviously $k_0 < \varkappa$.

For simplicity, we shall now assume $d_1 \geqq d_2$. In the other case, we should proceed in an analogous way (only with the difference that points with arguments $j k_0 \varphi$ for growing j move along the unit circle in the opposite sense).

Let $k_1^* = [2\pi/\delta]$, i.e. we have

$$\widehat{k_1^* \delta_0} > \widehat{(k_1^* + 1)\delta_0}.$$

At the same time

$$k_1^* < \varkappa - 1.$$

Let $\delta_1 = \min(d_1', d_2')$ where

$$d_1' = 2\pi - \widehat{k_1^* \delta_0}, \quad d_2' = \widehat{(k_1^* + 1)\delta_0}.$$

Obviously

$$\delta_1 \leqq \delta_0/2.$$

Let us define k_1 by

$$k_1 = \begin{cases} k_1^* & \text{if} \quad d_1' < d_2' \\ k_1^* + 1 & \text{if} \quad d_1' \geqq d_2', \end{cases}$$

so that we have $k_1 < \varkappa$.

If we now replace δ_0, k_0 by δ_1, k_1, respectively, and repeat the process, we can construct sequences $\{\delta_n\}$, $\{k_n\}$ such that

$$\begin{gathered} \delta_n \leqq \delta_0/2^n < \pi/2^n, \\ k_n < \varkappa \end{gathered} \tag{23}$$

for $n = 1, 2, \ldots$. However, the construction of δ_n shows that

$$\delta_n = |\widetilde{K_n \varphi}|$$

where $K_n = k_1 k_2 \dots k_n$. Hence for all n we have

$$K_n < \varkappa^n. \tag{24}$$

Combining (23) and (21), we see that $\delta_n \leqq b$ holds for $n \geqq N$. This implies the inequality

$$|\widehat{K_N \varphi}| \leqq b.$$

By (24), we have $K_N \leqq \varkappa^N$, where $\varkappa^N$ is independent of the choice of φ. This proves that no number ψ_m can enjoy the property (iii), and hence the statements (i) and (ii) are contradictory.

Consequently, there exists a sequence of natural numbers $\{L_n\}$ such that

$$\lim_{n \to \infty} \cos(L_n \psi_\nu) = 1$$

for $\nu = 1, 2, \dots, m$.

This completes the proof of the theorem.

Acknowledgement. The author is grateful to Dr. I. Babuška for giving the motivation for this research, the greater part of which was done at the Mathematics Institute of the Czechoslovak Academy of Sciences in Prague.

References

1. Bauer, F. L.: Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme. ZAMM **46**, 409—421 (1966).
2. Bieberbach, L.: Einführung in die Funktionentheorie (3rd ed.). Stuttgart: Teubner 1959.
3. Businger, P. A.: Matrices which can be optimally scaled. Numer. Math. **12**, 346—348 (1968).
4. Forsythe, G., Moler, C. B.: Computer solution of linear algebraic systems. Englewood Cliffs: Prentice Hall, Inc. 1967.
5. Gaier, D.: Konstruktive Methoden der konformen Abbildung. Berlin-Göttingen-Heidelberg: Springer 1964.
6. Kantorovič, L. V., Krylov, V. I.: Približennyje metody vyšego analiza (4th ed.). Moscow-Leningrad: Gostechizdat 1952.
7. Markuševič, A. I.: Teorija analitičeskich funkcij. Moscow-Leningrad: Gostechizdat 1950.

Dr. H. Švecová
Fachgruppe Computer-Wissenschaften
Eidg. Technische Hochschule
Leonhardstr. 33
CH-8006 Zürich

Hinweise für die Autoren

Die Autoren werden freundlichst gebeten, bei der Abfassung ihrer Manuskripte folgende Punkte zu beachten:

Dem Text ist ein englisches *Summary* voranzustellen, das für Kleindruck zu kennzeichnen ist.

Die verwendeten Symbole müssen so klar bezeichnet sein, daß auch beim Laien kein Zweifel über Stellung und Deutung auftreten kann. *Alle Formeln sind möglichst mit der Maschine zu schreiben*, wobei darauf zu achten ist, daß Indizes und Exponenten trotz des fehlenden Größenunterschiedes genau als solche zu erkennen sind. Anderenfalls müssen sie in geringerer Größe mit der Hand eingetragen werden. Besondere Lettern (griechische, gotische, Script) sind durch farbige Unterstreichung zu kennzeichnen. Zur Vermeidung von Verwechslungen wird empfohlen, *griechische Buchstaben rot, gotische Buchstaben blau und Scriptbuchstaben grün* zu unterstreichen. Es wird gebeten, auch gotische und Scriptbuchstaben mit der Maschine als lateinische Buchstaben zu schreiben und allein durch die farbige Unterstreichung zu kennzeichnen. Sofern dennoch handgeschriebene Buchstaben vorkommen, unterscheide man (auch bei lateinischen Buchstaben) große und kleine Buchstaben; *große Buchstaben sollten zweimal, kleine einmal unterstrichen werden.* Dies ist besonders wichtig bei *c, C*; *k, K*; *o, O*; *p, P*; *s, S*; *u, U*; *v, V*; *w, W*; *x, X*; *z, Z*. Besonders sorgfältig beachte man die Schreibweise bei handschriftlichen Buchstaben, wenn gleichzeitig *e* und *l* oder *u* und *n* oder *n* und *r* oder *o* und 0 als mathematische Bezeichnungen auftreten. Auch gleichzeitig auftretende *v* und ν sowie $\in$ und ε geben zu Verwechslungen Anlaß.

Sämtliche Buchstaben in Formeln, Einzelbuchstaben im Text sowie unterstrichene Textstellen werden automatisch kursiv gesetzt. Daher müssen z. B. in Formeln auftretende Abkürzungen, die in Antiqua (d. h. der üblichen Textschrift) gesetzt werden sollen, besonders gekennzeichnet werden (möglichst durch gelbe Unterstreichung). Fettzusetzende Buchstaben sind durch braune Unterstreichung zu kennzeichnen. Man vermeide in Formeln und bei mathematischen Symbolen das Unterstreichen mit Tinte oder Schreibmaschine; dies würde zwangsläufig als zum mathematischen Sinn gehörig interpretiert und daher mitgesetzt werden. Besondere Schwierigkeiten entstehen dadurch, daß Schreibmaschinen im allgemeinen keine Unterschiede zwischen 0 (Null) und O (Buchstabe) sowie häufig auch zwischen 1 (Eins) und l (Buchstabe) kennen. Hier sind unterscheidende Kennzeichnungen unbedingt erforderlich. Die arabische Ziffer 1 schreibe man deutlich als 1 und nicht einfach als senkrechten Strich. Das ist besonders verhängnisvoll, wenn die 1 als oberer Index auftritt. In englischsprachigen Manuskripten schließe man, wenn der kleine lateinische Buchstabe *a* als mathematische Bezeichnung vorkommt, die Verwechslung mit dem unbestimmten Artikel aus.

Durch derartige Mißverständnisse sind vielfach große Korrekturkosten entstanden. In den Korrektur-Abzügen sollen nur Satzfehler verbessert, jedoch keine inhaltlichen oder stilistischen Änderungen vorgenommen werden. Nachträgliche (vom Manuskript abweichende) Korrekturen müssen den Autoren in Rechnung gestellt werden.

Sämtliche zu einer Arbeit gehörende (sowohl photographische als auch Kurven und schematische) **Figuren** sind durchzunumerieren. Sie werden getrennt vom Text auf gesonderten Blättern erbeten.

Als *Vorlagen* werden Original-Kurven oder saubere, in klarem Schwarz und in einheitlicher Linienstärke angelegte Tuschzeichnungen erbeten. Die *Beschriftung* sämtlicher Figuren mit Buchstaben/Worten/Ziffern *erfolgt durch den Verlag*. Die Hinweise hierfür dürfen deshalb nicht in die Figur selbst eingezeichnet werden, sondern werden auf einem über die Vorlage geklebten transparenten Deckblatt erbeten. Wünsche des Autors hinsichtlich des linearen Verkleinerungs- oder Vergrößerungs-Maßstabes sollten auf der Rückseite der Vorlage mit weichem Blei vermerkt werden. Hierbei, insbesondere bei Bildgruppen, ist der zur Verfügung stehende Satzspiegel (122 × 194 mm) zu berücksichtigen.

Literatur-Zitate sind als geschlossenes, numeriertes Verzeichnis am Ende der Arbeit nach den Verfassernamen *alphabetisch* anzuordnen. Bei Zeitschriften-Zitaten sind folgende Angaben unerläßlich: Initialen und Namen sämtlicher Autoren, vollständiger Titel der Arbeit, Zeitschriftentitel, Band-, Seiten- und Jahreszahl. Bücher werden mit Autorennamen, vollem Titel, Auflage, Ort, Verlag und Jahr zitiert.

Fußnoten, die nicht zum Beitragskopf gehören, sind durchzunumerieren.

Der **Kolumnentitel** (Seitenüberschrift) darf 76 Buchstaben einschließlich Wortzwischenräumen nicht überschreiten. Bei umfangreicheren Beitragstiteln wird der Autor gebeten, eine entsprechende Kurzfassung auf der ersten Manuskriptseite anzugeben.

Jedem Manuskript sind auf gesondertem Blatt „Anweisungen für den Satz" beizugeben, auf dem die benutzten Kennzeichnungen sowie verwendete besondere Symbole erklärt werden.

Bei Einsendung eines Manuskriptes *(in doppelter Ausfertigung)* ist die **Adresse des Autors** genau anzugeben. *Wechselt* der Autor bis zum Erscheinen seiner Arbeit seine *Anschrift*, so bitten wir in seinem eigenen Interesse dringend *um sofortige Benachrichtigung.*

Directions for Authors

Authors are asked to kindly observe the following points in the preparation of their manuscripts for publication:

All papers should be preceded by a short *summary* in English which should be marked for small print.

Formulae should be written so that no uncertainty can arise as to the meaning and position of signs and characters, even if the reader is not an expert. *Formulae should be type written* and special care taken that indices and exponents can be recognized as such even if they show no difference in size. Otherwise they are to be written in a smaller size by hand. Special type (Greek, Gothic, Script) must be indicated by underlining in colour. In order to avoid confusions the underlining of *Greek characters in red, Gothic in blue and Script in green* is recommended. It is advisable to type Roman characters also for Gothic and Script and to distinguish them simply by underlining in colour. If, however, hand-written characters have been used, small and capital letters should be clearly distinguished *by underlining capital letters twice and small letters once.* This also applies to Roman characters written by hand such as *c, C; k, K; o, O; p, P; s, S; u, U; v, V; w, W; x, X; z, Z.* Special care should be taken to show the difference between *e* and *l* or *u* and *n* or *n* and *r* or *o* and 0, wherever they appear simultaneously as mathematical symbols written by hand. The characters *v* and ν as well as $\in$ and ε are also liable to cause confusion when they appear together.

All letters contained in formulae, as well as single letters in the text are automatically composed in italics and therefore require no underlining. On the other hand abbreviations in Roman type (the type normally used for the text) that appear in formulae, must be specially marked (by underlining in yellow, if possible). Bold type should be specially marked by underlining in brown. Underlining of formulae and mathematical signs in ink or by means of a typewriter should be avoided as this would inevitably lead to inclusion of the underlining in the composition as part of the mathematical symbol. The fact that typewriters usually show no difference between the figure 0 (zero) and the letter O, nor, frequently, between 1 (one) and the letter l, leads to considerable confusion. Here it is absolutely essential that each letter and each numeral be individually distinguished. The numeral 1 should be written clearly as such (i.e. with a preceding hook) and not merely as a downward stroke. The latter is particularly disastrous where 1 appears as superscript. In manuscripts written in English, in which the small (Roman) letter *a* is employed as a mathematical symbol, care must be taken to avoid any confusion with the indefinite article.

Such misunderstandings have frequently led to a considerable increase in the cost of proof. Corrections in the proofs should be restricted to typographical errors; changes as to grammatical and stylistic deficiences are to be avoided. Expenses accruing from such additional corrections will be charged to the authors.

All **figures** including graphs are to be numbered consecutively and should be submitted on separate sheets.

Line drawings and *graphs* should be drawn with indian ink in clean, uniform lines on smooth white paper or Bristol board. The *labelling* of all figures with letters, words, numerals *is done by the publisher.* Therefore lettering must not be placed on the figure but instead on a cover sheet of transparent paper. The author may indicate by means of a soft pencil on the reverse side of the illustrations instructions regarding the desired linear reduction or magnification. It must be emphasized that the maximum area available for the reproduction of a figure (or an array) is 122×194 mm.

The reference list, to be placed at the end of the paper, should be *in alphabetical order* of the names of the first author. For journal articles the following information should be provided: names and initials of all authors, complete title of paper, name of Journal, number of volume, first and last pages, and year of publication. Books are to be cited by listing the author (s), full title, edition, place of publication, publisher, and year.

Footnotes which do not belong to the heading of the paper should be numbered consecutively.

The running head (condensed title) should not exceed 76 letters including spaces. If the title of the paper is very long the author is requested to kindly state an adequate condensation of it on the first page of the manuscript.

Each manuscript should be accompanied by a separate sheet bearing "Instructions for the compositor" which explains the meaning of marks and other symbols used.

Manuscripts should be submitted in duplicate.

The address of the author must be clearly indicated. If the author *changes his address* before his paper is published he is requested to notify us *immediately.*

Numerische Mathematik

Band 14 · (Schluß-) Heft 5 · 1970

Inhalt

Indexed in Current Contents

Verantwortlich für den Anzeigenteil: Edgar Seidler, 1 Berlin 33, Heidelberger Platz 3
Springer-Verlag, Berlin · Heidelberg · New York
Druck der Universitätsdruckerei H. Stürtz AG, Würzburg